U.S. DEPARTMENT OF COMMERCE
National Technical Information Service
PB-292 581

Design of Field Experiments to Determine the Ecological Effects of Petroleum in Intertidal Ecosystems

Resource Management Associates, Lafayette, CA

Prepared for

National Oceanic and Atmospheric Administration, Boulder, CO

Dec 78

NOAA

National Oceanic and Atmospheric Administration
Boulder CO 80303

EPA-600/7-78-231
December 1978

EPA

United States
Environmental Protection
Agency

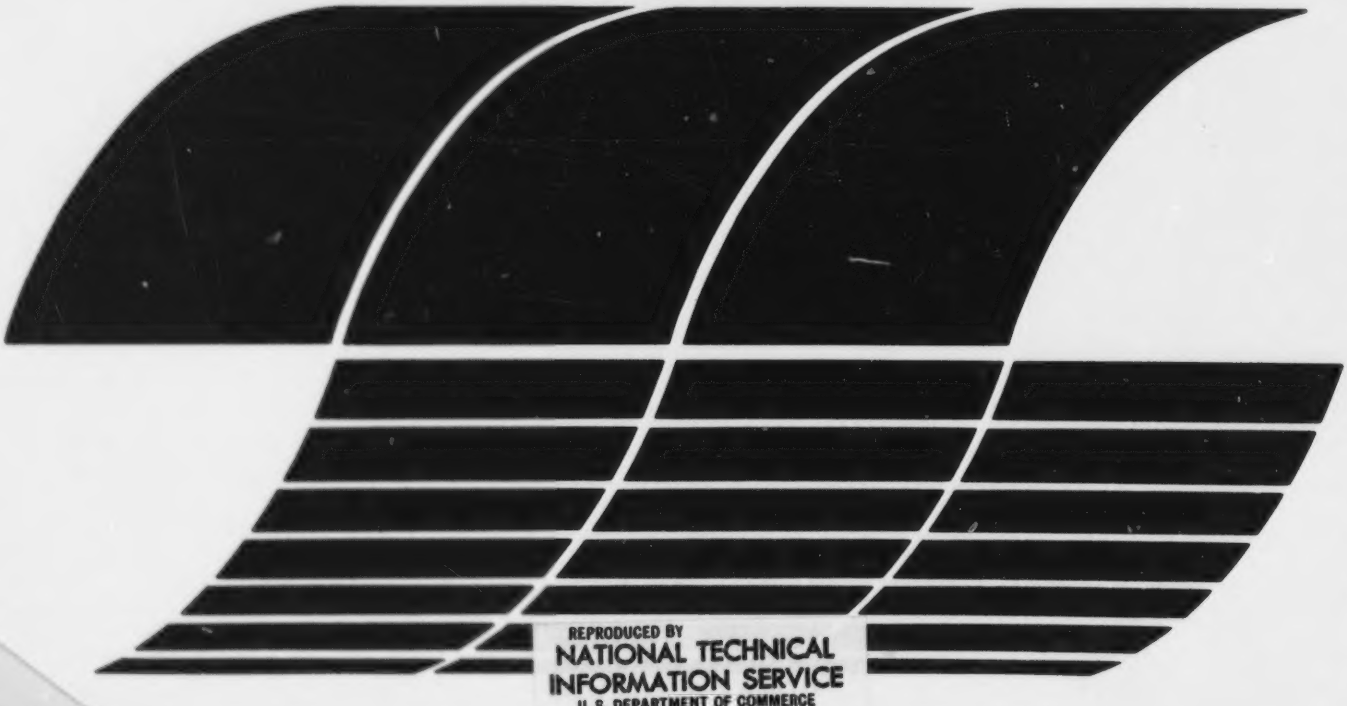
Office of Energy, Minerals, and
Industry
Washington DC 20460

PB 292581

Research and Development

Design of Field Experiments to Determine the Ecological Effects of Petroleum in Intertidal Ecosystems

Interagency Energy/Environment R&D Program Report



REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the INTERAGENCY ENERGY-ENVIRONMENT RESEARCH AND DEVELOPMENT series. Reports in this series result from the effort funded under the 17-agency Federal Energy/Environment Research and Development Program. These studies relate to EPA's mission to protect the public health and welfare from adverse effects of pollutants associated with energy systems. The goal of the Program is to assure the rapid development of domestic energy supplies in an environmentally-compatible manner by providing the necessary environmental data and control technology. Investigations include analyses of the transport of energy-related pollutants and their health and ecological effects; assessments of, and development of, control technologies for energy systems; and integrated assessments of a wide range of energy-related environmental issues.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO.
EPA-600/7-78-231

2.

3. RECIPIENT'S ACCESSION NO.
PB292581

4. TITLE AND SUBTITLE
Design of Field Experiments to Determine the Ecological Effects of Petroleum in Intertidal Ecosystems

5. REPORT DATE
December 1978 issuing date

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Stephen F. Moore and Dennis B. McLaughlin
Resource Management Associates
3706 Mt Diablo Blvd, Suite 200, Lafayette CA 94541

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
National Oceanic and Atmospheric Administration
Environmental Research Laboratories
Boulder CO 80303

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
03-6-022-35258

12. SPONSORING AGENCY NAME AND ADDRESS
U.S. Environmental Protection Agency
Office of Research & Development
Office of Energy, Minerals & Industry
Washington, D.C. 20460

13. TYPE OF REPORT AND PERIOD COVERED
FINAL

14. SPONSORING AGENCY CODE
EPA/600/17

15. SUPPLEMENTARY NOTES
This project is part of the EPA-planned and coordinated Federal Interagency Energy/Environment R&D Program.

16. ABSTRACT

Is it possible to design field experiments that will yield ecologically and statistically significant information about how oil affects intertidal ecosystems? What classes of experimental design and technical approach are most likely to generate optimal information on these effects? In order to improve the usefulness of field experiments to prediction and assessment of impacts of oil spills on marine environments, this report addresses the foregoing questions as they apply to rocky intertidal habitats characteristic of the Gulf of Alaska. The report discusses problems of experimental design in the ecosystems and presents statistical approaches for dealing with the problems. Examples are provided using data on rocky shore habitats at Zaikof Bay, Alaska. The levels of variability exhibited by these data indicate that realistic experiments can be designed to study the effects of oil. Traditional experimental design methods of factor selection, randomization, blocking and performance evaluation are directly relevant to intertidal oil experiments. The experimental effort, or cost, required to document a particular ecological effect can be systematically evaluated as a function of experimental design variables and can be estimated using data previously available.

17. (Circle One or More) KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Ecology Environments Earth Atmosphere Environmental Engineering Geography Hydrology, Limnology Biochemistry Earth Hydrosphere Combustion Refining Energy Conversion Physical Chemistry Materials Handling Inorganic Chemistry Organic Chemistry Chemical Engineering Other: Experimental Design	[illegible]	6F 8A 8F 8H 10A 10B 7B 7C 13B

18. DISTRIBUTION STATEMENT RELEASE UNLIMITED	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21. NO. OF PAGES 196
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE PC A09/MF A01

EPA Form 2220-1 (9-73)

☆ U.S. GOVERNMENT PRINTING OFFICE:1979—677-072 / 1229

EPA-600/7-78-231
December 1978

DESIGN OF FIELD EXPERIMENTS TO DETERMINE THE ECOLOGICAL EFFECTS OF PETROLEUM IN INTERTIDAL ECOSYSTEMS

by

Stephen F. Moore
Dennis B. McLaughlin
Resource Management Associates
3706 Mt. Diablo Blvd., Suite 200
Lafayette, CA 94549

Contract No. 03-6-022-35258

Project Officer: Douglas A. Wolfe

National Oceanic and Atmospheric Administration
Environmental Research Laboratories
Boulder, Colorado 80303

Office of Research and Development
U. S. Environmental Protection Agency
Washington, D. C. 20460

NOTICE

This work was supported jointly by the National Oceanic and Atmospheric Administration and the Environmental Protection Agency under the Federal Interagency Energy/Environment R&D Program. Since 1975, the Environmental Protection Agency had had the lead responsibility for the planning, coordination, and implementation of this program, which is funded at approximately $100 million per year and participated in by seventeen Federal departments and agencies. The research reported here represents one part of NOAA's Energy/Environment Project entitled "Fate and Effects of Petroleum Hydrocarbons and Toxic Metals in Marine Ecosystems and Organisms."

The National Oceanic and Atmospheric Administration and the Environmental Protection Agency do not approve, recommend, or endorse any proprietary product or proprietary material mentioned in this publication. No reference shall be made to the Environmental Research Laboratories or to this publication furnished by the Environmental Research Laboratories in any advertising or sales promotion which would indicate or imply that the Environmental Research Laboratories approve, recommend, or endorse any proprietary product or proprietary material mentioned herein, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of this Environmental Research Laboratories publication.

FOREWORD

As this Nation's effort towards energy self-sufficiency is carried forward, a considerable amount of environmental information will be required to permit energy developments to proceed with acceptable impact on environmental quality. Under the terms of an Interagency Agreement between the National Oceanic and Atmospheric Administration and the Environmental Protection Agency, NOAA is conducting studies on the Fate and Effects of Petroleum Hydrocarbons and Selected Toxic Metals in Selected Marine Organisms and Ecosystems. The overall objectives of the project are to study experimentally those specific processes controlling the distribution, transport and effects (physiological and ecological) of petroleum hydrocarbons and toxic metals in coastal marine ecosystems to facilitate the assessment of impacts of petroleum releases and to serve as the basis for developing regulatory measures for suitable protection of the marine environment. It is anticipated that carefully designed and conducted field experiments will play a major role in documenting the ecological effects of petroleum in marine systems. This report was viewed as a necessary first step in assessing the feasibility and most useful approaches for such field experiments.

Douglas A. Wolfe
Deputy Director, Outer Continental Shelf
Environmental Assessment Program
NOAA Environmental Research Laboratories

ACKNOWLEDGMENTS

We express a special thanks to Dr. Douglas A. Wolfe, Deputy Director, Outer Continental Shelf Environmental Assessment Program. His patience, support and careful review have been extraordinary throughout the year and a half we have engaged in this effort.

We also thank the following persons who participated in our workshop on the Design of Oil Spill Perturbation Field Experiments, December 13 and 14, 1976: Dr. Jack Anderson, Dr. Robert Clarke, Dr. Jeff Fujioka, Mr. Edward Long, Dr. Bill MacLeod, Dr. Dick Nyren, Dr. Robert Payne, Dr. Steve Zimmerman. We received many useful inputs at the workshop, which have contributed substantively to this final report.

Several persons received an earlier draft of this report. We have made many improvements in the final version due to their reviews. Our thanks for these reviews go to Dr. David Colby, Dr. Robert Paine, Dr. Lee Eberhardt, Dr. Ken Mann, Dr. Jack Anderson and Dr. J. Vanderhorst.

Our thanks to Dr. Jim Audet NODC OCSEAP Data Coordinator for providing us with magnetic tape copies of the Alaskan data sets used in our analysis.

Finally, we thank Ms. Lillian Orlob and Ms. Carol Hunter for the efficient and patient deciphering, typing and retyping of this report.

TABLE OF CONTENTS

PAGE

ACKNOWLEDGMENTS iv

1. PRELIMINARIES 1

1.1 PROBLEM MOTIVATION 1
1.2 OUR PERSPECTIVE AND BACKGROUND 2
1.3 SCOPE OF STUDY 3
1.4 THE EXPERIMENTAL DESIGN PROBLEM 4
1.5 SOME FEATURES OF ROCKY SHORES 8

2. EXPERIMENTAL DESIGN: ECOLOGICAL ASPECTS 11

2.1 INTRODUCTION 11
2.2 EXPERIENCE WITH ROCKY SHORE EXPERIMENTS 12
2.3 REPORTED EFFECTS OF OIL: A BASIS FOR CHOOSING TESTABLE HYPOTHESES 15
2.4 OIL AS AN EXPERIMENTAL FACTOR 23
2.5 CHOOSING EXPERIMENTAL UNITS AND OBSERVATIONS 26
2.6 SUMMARY OF ECOLOGICAL ASPECTS OF DESIGN 28

3. EXPERIMENTAL DESIGN: STATISTICAL ASPECTS 31

3.1 INTRODUCTION 31
3.2 FUNDAMENTAL CONCEPTS 35

3.2.1 Experimental Models 35
3.2.2 Estimation of Treatment Effects and Population Parameters 45
3.2.3 Detection of Significant Treatment Effects 54

3.3 FACTORS INFLUENCING THE PERFORMANCE AND FEASIBILITY OF EXPERIMENTAL DESIGNS 64

3.3.1 General Design Considerations 64
3.3.2 Qualitative Aspects of Experimental Design--Factor Selection, Randomization, Blocking, and Quadrat Placement 68
3.3.3 Quantitative Aspects of Experimental Design--Decision Thresholds, Design Parameters, and Feasibility Assessment 80

3.4 SUMMARY OF STATISTICAL DESIGN CRITERIA 101

4. EXAMPLE DESIGNS OF ROCKY SHORE OIL EXPERIMENTS FOR ZAIKOF BAY, ALASKA 106

4.1 INTRODUCTION 106
4.2 EXPERIMENTAL SITE DESCRIPTION - ZAIKOF BAY, ALASKA 107

TABLE OF CONTENTS
(continued)

PAGE

4.2.1 Existing Data from Zaikof Bay 108
4.2.2 Statistical Analysis of the Zaikof Bay Data 113

4.3 EXPERIMENTAL OBJECTIVES AND STRATEGY 115
4.4 DESIGNS 118

4.4.1 Preliminary Surveys 119
4.4.2 Initial Manipulation Experiments 120
4.4.3 Oil Feasibility Experiments 138
4.4.4 Oil Impact Experiments 141
4.4.5 Recovery Experiments 153

4.5 IMPLICATIONS AND GENERALIZATIONS 153

4.5.1 Objectives, Factors and Treatments 154
4.5.2 Layouts and Number of Samples 155

5. CONCLUSIONS AND RECOMMENDATIONS 157

5.1 CONCLUSIONS 157
5.2 RECOMMENDATIONS 160

REFERENCES 162

APPENDIX A ESTIMATION AND HYPOTHESIS TESTING WITH MIXED EFFECT MODELS 166

APPENDIX B ALTERNATIVE TECHNIQUES FOR ASSIGNING TREATMENTS TO SAMPLING QUADRATS 169

LIST OF FIGURES

PAGE

Figure 1-1 Principal Trophic Relationships in Rocky/Cobble Intertidal Habitat 10

Figure 3-1 An Example of the Problems Encountered in Non-Stratified Sample Surveys 38

Figure 3-2 Relation of Residual Error, Observation, and Effect Estimate Probability Density Functions for a Fixed Effect Model 47

Figure 3-3 Probability Densities of the F Statistic Used To Test the General Linear Hypothesis 58

Figure 3-4 Number of Treatments Required in a Single Factorial Replicate vs. Number of Experimental Factors 69

Figure 3-5 Example of the Application of Randomization to an Experimental Design 72

Figure 3-6 Example of Quadrat Location Problems Posed by Petroleum Treatments 77

Figure 3-7 Design Parameters Affecting the Risk Associated with an F-test of the Linear Hypothesis H_A 83

Figure 3-8 Plot of Type II vs. Type I Error Probabilities for a Particular Experimental Layout (adapted from Figure 4-6b) 86

Figure 3-9 Plot of Normalized Risk vs. Significance Level for Various Relative Cost Alternatives 87

Figure 3-10 Feasibility Evaluation for a 4x2x2x2 Layout 100

Figure 4-1 F-test Power vs. Normalized "predator effect" (Δ/σ_e) with Different Numbers of Replicates in a 4×2^3 Complete Block Layout 126

Figure 4-2a Detectable Change as a Percentage of the Mean for *B. Glandula* vs. Number of Replicates for Different Powers in a 4×2^3 Complete Block Layout 128

Figure 4-2b Detectable Change as a Percentage of the Mean for *Rhodymenia palmata* vs. Number of Replicates for Different Powers in a 4×2^3 Complete Block Layout 129

LIST OF FIGURES

(continued)

PAGE

Figure 4-2c Detectable Change as a Percentage of the Mean for *Odonthalis dentata* vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout 129

Figure 4-2d Detectable Change as a Percentage of the Mean for *Mytilus edulis* vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout 129

Figure 4-2e Detectable Change as a Percentage of the Mean for *Nucella lamellosa* vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout 129

Figure 4-3 Delta vs. Replicates for Different Values of σ_e in 4 x 2^3 Complete Block Experimental Layout 131

Figure 4-4 Detectable Change in *B. glandula* as a Percentage of the Mean vs. Replicates of a 4 x 2^3 Complete Block Layout for Power of .85 and Various Levels of Significance (α) 132

Figure 4-5 Power vs. Alpha for Detecting Different Levels of Change in *B. glandula* with Four Replicates of a 4 x 2^3 Complete Block Layout 134

Figure 4-6a Power vs. Alpha for Different Values of Δ/σ_e with Four Replicates of 4 x 2^3 Complete Block Layout 135

Figure 4-6b Power vs. Alpha for Different Values of Δ/σ_e with Twenty Replicates of 4 x 2^3 Complete Block Layouts 135

Figure 4-7 Marginal Sensitivity of Residual Error Standard Deviation to Replicates for Power .85 and Different Levels of Detectable Change in a 4 x 2^3 Complete Block Experimental Layout 137

Figure 4-8 Conceptual Representation of the "Smearing" Problem Associated With Using Oil as a Factor (Scale is Arbitrary) 139

Figure 4-9 F-test Power vs. Δ/σ_e for Testing Interaction Hypothesis in a 2 x 2 Complete Block Layout 147

Figure 4-10 Detectable Percent Change in Mean of *B. glandula* Due to Interaction Effect vs. Replicates for Different Levels of Power in 2 x 2 Complete Block Layouts 148

LIST OF FIGURES
(continued)

PAGE

Figure 4-11 Detectable Change Due to Interaction vs. Replicates at Different Values of σ_e in 2 x 2 Complete Block Layout 149

Figure 4-12a Power vs. Alpha for Different Levels of Δ/σ_e with Eight Replicates of 2 x 2 Complete Block Experiment 151

Figure 4-12b Power vs. Alpha for Different Levels of Δ/σ_e with Forty Replicates of 2 x 2 Complete Block Experiment 151

Figure 4-13 Marginal Sensitivity of σ_e to Replicates at Different Levels of Change for 2 x 2 Complete Block Experiments 152

Figure B-1 Examples of Complete and Incomplete Block Layouts 171

Figure B-2 Example of Confounding for a 2x2x2 Layout 174

LIST OF TABLES

PAGE

Table 3-1 Factorial Treatments for a 2 Factor Controlled Experiment 40

Table 3-2 Degree of Freedom Allocations for the General Complete Block Factorial Model 90

Table 3-3 Analysis of Variance Table for the General Complete Block Factorial Experiment 97

Table 4-1 Selected Species List for Zaikof Bay 109

Table 4-2 Taxonomic Groups Not Included in Selected Species List for Zaikof Bay 110

Table 4-3 Zaikof Bay Data Set 112

Table 4-4 Statistics of Selected Species at Zaikof Bay 114

Table 4-5 Residual Error Variance for Selected Species at Zaikof Bay Using One-Way ANOVA Model on Tidal Elevation 116

Table 4-6 Power of F-Test on Main Effects for 4x2x2x2 Experimental Layout in the Case of $\Delta/\sigma_e = 1, \alpha = .05$ 125

Table A-1 Distributions of Mixed-Effect Test Statistics 167

Table B-1 Comparison of Degree of Freedom Allocations for Unconfounded and Completely Confounded Designs 177

1. PRELIMINARIES

1.1 PROBLEM MOTIVATION

Is it possible to design a field experiment that will yield ecologically and statistically significant information about how oil affects intertidal ecosystems? What classes of experimental design and technical approach are most likely to generate optimal information on these effects? These questions arise in the process of gathering information to enable prediction and assessment of impacts of oil and gas developments on marine environments. Information is needed to support decisions regarding the exploration and development of offshore petroleum resources; and regulations must be established regarding leasing, operating and monitoring which permit resource developments with maximum environmental protection. In order to generate this information, the National Oceanic and Atmospheric Administration (NOAA), with support from the Environmental Protection Agency (EPA), is conducting studies on the Fate and Effects of Petroleum Hydrocarbons and Selected Toxic Metals in Marine Organisms and Ecosystems. This report describes the results of one small part of this program aimed at answering the questions raised above.

Field perturbation experiments intended to test and demonstrate the effects of oil spills on intertidal habitats are one of several research approaches which ultimately support environmental policy and regulatory objectives. By their very nature these objectives raise "what if" questions about petroleum resource developments. Comprehensive answers to these questions must be based on generalizations derived from observations made in field and laboratory studies. However, such generalizations are risky unless causes of observed phenomena are understood. Consequently, an experimental oil spill should be designed particularly to help improve our understanding of *causes* of observed phenomena--a program which will support the generalizations needed for policy decisions.

1.2 OUR PERSPECTIVE AND BACKGROUND

At the outset we point out features of our background and perspective which we believe influence the results of our analysis. To begin with, we are neither biologists nor field ecologists and our lack of such experience no doubt lends a certain naiveté to some of our descriptions and examples. Our point of view reflects our experience in "system science"--modeling, estimation, experiment design, systems ecology and so on. We believe that the task of *designing* an experiment is in many ways an engineering problem, requiring assumptions, predictions and hypotheses about as yet unobserved events. Yet, this is only part of the problem and an understanding of intertidal ecosystems and the behavior of oil is also necessary. Where our own experience limits us in these areas, we have sought input from knowledgeable persons. Specifically, we conducted a small workshop as part of our research effort. The input received during and after the workshop has been most valuable. Of course, we are solely responsible for errors of fact and judgment which remain in the report itself.

Some other aspects of our perspective warrant noting. We have attempted to prepare a report which is responsive to the needs of the field investigator. Statistical design methods have not often appeared particularly useful in field studies. Problems which frequently arise include:

- Requirements for designs which require unrealistically large numbers of samples.
- Designs which yield data that do not verify ecological differences which are visually obvious.
- Designs which cannot be implemented due to variability encountered at the experimental site.

In our opinion, these problems motivate the need for design criteria and methods which are flexible and sequential. The field investigator must also be willing to exercise considerable subjective judgment and not rely only on traditional methods for developing experimental designs.

Finally, we have found this research effort to be almost overwhelming at times. Much of the time we have operated at the boundaries of our understanding of the problem. There are so many pieces to the problem, and so few that we know anything about; and even in the subject areas with which we are familiar, the existing literature is massive. We make no apologies nor excuses--we are simply venting our frustration.

1.3 SCOPE OF STUDY

In order to make this study manageable, several constraints were placed on the scope of our analysis. First, the western Gulf of Alaska and Puget Sound areas form the geographical regions of interest. Secondly, we restrict ourselves to rocky intertidal habitats. To the extent that we invoke examples and representations of specific community patterns and interactions, we have in mind rocky intertidal shores typical of this restricted geographical area. In many ways, however, these constraints do not limit the scope of the results of our analysis; the criteria for experimental design could be applied equally to other habitat types in other geographical areas.

Pacific rocky intertidal shores have several characteristics which make them a desirable focus for analysis. By lineal or areal measure they are the dominant intertidal habitat for the Northwest and Alaskan Gulf coast. Rocky shores are relatively well-studied, in general. Many surveys and experimental studies of rocky habitats are reported in the literature. (See Section 2.2 for references). Several accidental oil spills in rocky areas have been observed, so we have some knowledge of how oil can affect these systems. (See Section 2.3 for references.) More specifically, the rocky shores in and near Puget Sound have been studied for over 10-15 years by R. T. Paine and his co-workers at the University of Washington. As Connell (1972) describes, rocky shores have several desirable features for experimental analysis. Most importantly, the populations found in rocky shore habitats are amenable to experimental manipulation in the field, thereby allowing direct measurement of dynamic relationships between organisms. In addition, rocky intertidal habitats provide an accessible and observable environment in which to perform field studies to investigate ecosystem responses to environmental changes.

Our focus is the design of field studies which elucidate and quantify causal relationships. With respect to oil, we are primarily concerned with explaining less-obvious, subtle effects as well as visually obvious mortality due to smothering by massive coating of oil in an intertidal area. Our interest is in determining the causes of effects, not only observing effects. In many cases, these subtle effects might occur over the long-term, which would require experiments lasting over periods of years simply for the effects to exhibit themselves. In such cases other sources of variability or ecological interaction may mask the hypothesized effect such that demonstration of causality becomes a statistical improbability. Such long-term objectives and experiments may furthermore, be quite expensive, thereby rendering the experimental spill approach infeasible for such objectives.

Whether to implement a "controlled" oil spill involves important scientific and policy questions. We have restricted this report to the scientific feasibility of such a study. However, other important questions concerning site selection, potential environmental impacts, and permitting must also be addressed before this kind of experiment should be implemented. In our opinion, a scientific analysis and feasibility study should be part of the input to the policy decision. This study may serve both as a guide for experimental design of future field studies, and as a basis for policy decisions on proposed studies of similar intent.

1.4 THE EXPERIMENTAL DESIGN PROBLEM

It is convenient at this point to introduce a standard statistical terminology for the subject of experimental design. A number of the terms frequently used in subsequent chapters are defined in a general way in the list presented below. In some cases, more detailed or quantitative definitions are provided when the relevant concepts have been discussed more thoroughly.

Controlled Experiment - As we use the term, a controlled experiment is a study designed to identify causal relationships between observed *dependent variables* (e.g., species abundance) and various *independent variables* (e.g., tidal elevations, substrate composition, time of year, etc.)

thought to affect these observations. The word *controlled* in this term refers to the fact that some or all of the independent variables may be manipulated or constrained by the experimenter.

Independent and Dependent Variables - These terms are used to ascribe cause and effect to field variables investigated in a controlled experiment. Independent variables represent the causes responsible for effects observed in the dependent variables. Although causality can rarely be established beyond a doubt, careful experimental design does permit an investigator to infer it with reasonable likelihood.

Sample Surveys - This term refers to a field study intended to estimate the mean of a particular population, without regard for causal factors.

Analytical Surveys - This type of survey is concerned with the comparison of the means of two or more populations. The population means of interest are often species abundances (or similar ecological measures) at two sites or at two times.

Sample Stratification - This is a procedure sometimes used to provide better mean estimates in survey studies. It is based on the concept that the population of interest is arranged, for various reasons, in distinct strata or categories which account partially for observed variability.

Experimental Factor - This is a term assigned to an independent variable being investigated in a controlled experiment. An experiment designed to study such factors is called a *factorial experiment*.

Factor Level - The levels of a factor are used to classify or measure the factor for experimental purposes. If the factor is *quantitative*, its levels refer to particular values or ranges of values (e.g., tidal elevation at a site may be divided into four levels corresponding to 0-1, 1-2, 2-4, and 4-6 meters above mean lower low water). If the factor is *qualitative*, its levels refer to several mutually exclusive conditions (e.g., limpet manipulation in a controlled experiment may be divided into two levels corresponding to limpet exclusion and non-exclusion).

Experimental Unit - In our application, an experimental unit is a small uniform area of the rocky shore site which forms the basic spatial observation unit of the experiment. The term *sampling quadrat* is used more or less synonymously.

Experimental Treatments - In factorial experiments, a treatment is a particular combination of factor levels applied to an experimental unit (e.g., limpet exclusion in the 1-2 meter tidal range is a distinct treatment in an experiment having four tidal elevation levels and two limpet manipulation levels).

Experimental Model - A mathematical description of an experiment which relates dependent variable observations to the various treatments applied to experimental units.

Treatment Effect - A treatment effect is the portion of a dependent variable observation attributed by the experimental model to a particular treatment.

Residual Error - Residual error is defined here as the portion of observation variability which cannot be explained by (attributed to) experimental factors.

Nuisance Variable - A nuisance variable is any potentially important independent variable which is not included as an experimental factor in a controlled experiment.

Experimental Block - A block is a set of experimental units or quadrats grouped together for the purpose of minimizing the residual error associated with a controlled experiment.

Randomization - Randomization is the process of allocating factorial treatments randomly to experimental units within a particular block. This is done to minimize the possibility of confusing the effects of nuisance variables or other extraneous phenomena with the effects of factorial treatments.

Replicate - A replicate is the complete set of samples (or experimental units) associated with all possible factorial treatments--one sample for each treatment.

In a relatively narrow analytical sense the objective of experimental design is to determine a set of cost-effective treatments, sample allocations and data analysis methods which provide the researcher with an acceptable probability of detecting a statistically significant effect when one exists. More broadly, Cox (1958) distinguishes two aspects of experimental planning:

- ° Qualitative Aspects of Experimental Design - The choice of observations (dependent variables), factors (independent variables), randomization and blocking procedures, and sampling quadrats to be used in a controlled experiment.
- ° Quantitative Aspects of Experimental Design - The specification of decision thresholds and/or penalty costs for computing decision risk, the selection of the number of treatment levels and replicates to be used, and the estimation of residual error values to be used in evaluating experimental feasibility.

The qualitative aspects of experimental design involve statistical concepts but are primarily dependent on the investigator's objectives and one's understanding of the ecological system under study. The judgments and skills required here are quite application-dependent. The quantitative aspects of experimental design are more closely related to the traditional methods of mathematical statistics. The techniques required here are usually general and relatively independent of the particular system and problem of interest.

We examine the application-dependent aspects of experimental design in the intertidal environment in Chapter 2 and deal with more general statistical and mathematical questions in Chapter 3. This division is somewhat arbitrary and is made primarily for convenience of presentation. In an actual application, ecological and statistical design issues are closely related. This is demonstrated in Chapter 4 where we present a synthesis of the methods and design criteria developed in Chapters 2 and 3.

1.5 SOME FEATURES OF ROCKY SHORES

An extensive literature exists describing Pacific rocky shores. The classical reference on Pacific shores is Ricketts, et al (1968). Kozloff (1973) provides a useful guide to seashore life of Puget Sound and nearby areas. Lewis (1964) is another classical reference based on the rocky shores of the British Isles. An important review and synthesis of experimental studies is presented by Connell (1972). The features we describe below and assumptions we make about rocky shore communities are largely based on our interpretation of these references and our own limited excursions into rocky seashores.

Rocky shores are obviously recognizable. Surely there are transitional zones that pose classification problems, but we are not concerned with these. Our attention is focused on those areas which can be readily identified as rocky intertidal. These areas are characterized by dominant physical and biological features which we expect to find wherever rocky shores occur. Some physical features, such as wave exposure, topography, and orientation have a modifying effect from site to site.

Biologically, the characteristic feature of rocky shores is vertical zonation (Lewis, 1964). Although most hypotheses about the cause of vertical zonation emphasize tides, Connell (1972) convincingly demonstrates the importance of biological interactions. In addition to vertical zonation, patterns of species distribution also change in the horizontal, correlated with gradients in water movement and degree of turbulence.

These physical factors and characteristic vertical and horizontal distributions of organisms make rocky shores recognizable. We expect to find these features. We also expect to find a high degree of variability and small-scale patchiness in species abundance and distribution. Knowing the details of species distribution at one site tells us very little about exact species distribution at another. At any given location the specific topography, orientation, wave exposure, rock size distribution and rock type gives rise to the particular observed species composition, species abundance, and exact tidal location and extent of zonation.

In addition to these descriptive features, certain functional processes also characterize rocky shore ecosystems. Most important are characteristic patterns of predation and competition. Although the particular species involved in an interaction may vary among sites, Connell (1961) has shown that rock surface area is the limiting resource exploited by groups of organisms ("guilds"). Paine (1974) has shown that predation can be an essential process in the allocation of this resource among competing species. Other processes which play important roles in the community organization of rocky shores are recruitment, settlement and vertical migration. Figure 1-1 shows a food web typical of a rocky shore community.

The foregoing description of rocky intertidal shores summarizes general features which characterize these habitats. We leave the demanding task of describing details to the many existing references available. We reiterate that, although exact abundance and distribution of species varys among sites, there are certain important features common to rocky shore habitats and which allow us to make some assumptions and predictions about the structure and function of these communities. It will be evident that without such information, the design of experiments would be a hopeless task. The apparent complexity and variability of any natural rocky shore community are formidable hurdles to studying and understanding observed community interactions. However, these difficulties should not prevent us from formulating and testing more general hypotheses explaining observed phenomena. The point is succinctly made by Paine (1974):

> "It is precisely because the dynamic interaction is predictable that the rocky intertidal is a biologically rich and vibrant community, one showing few signs of the lack of organization and absense of important or visible processes that one would associate with an ecologically stringent environment continually disrupted by events of unpredictable timing, position or magnitude (p. 118)."

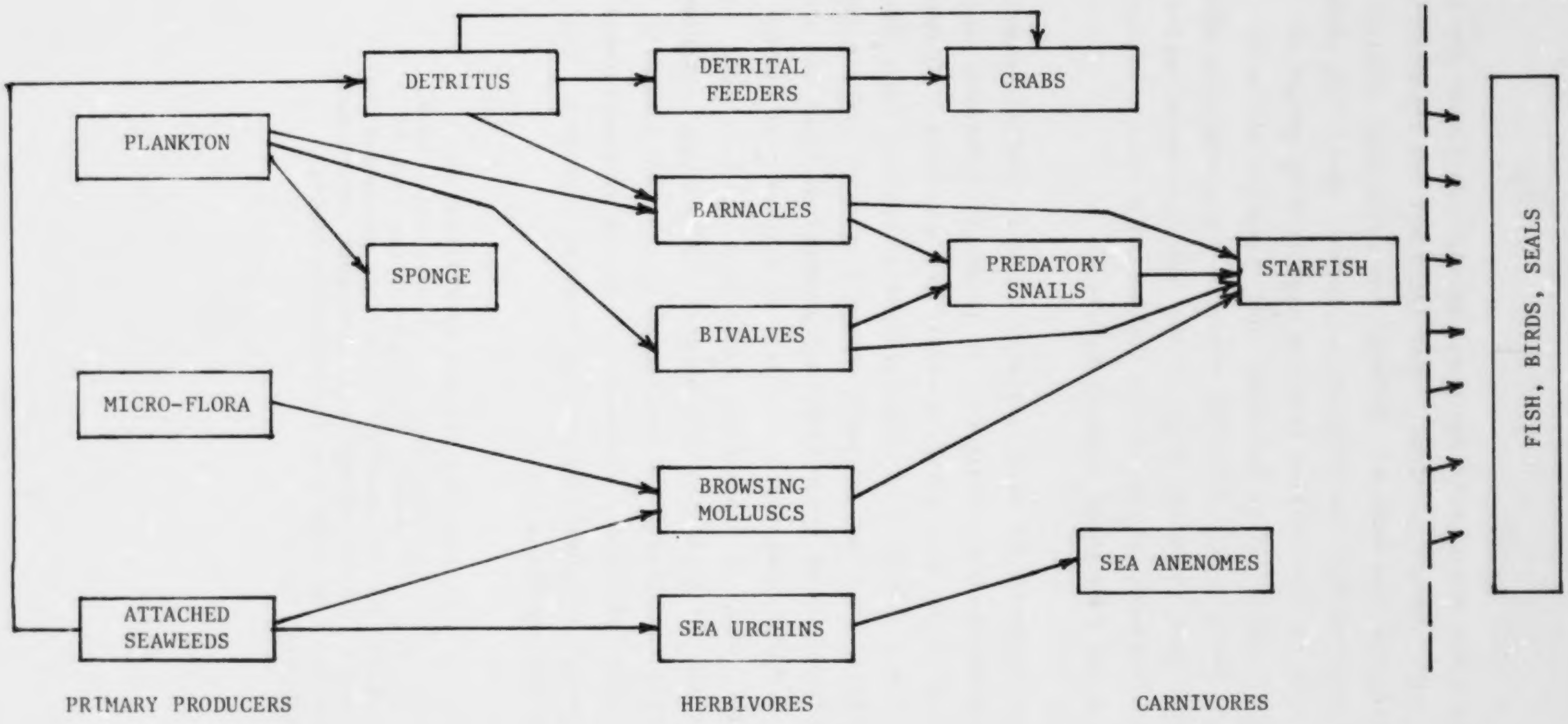

DETRITUS
DETRITAL FEEDERS
CRABS
PLANKTON
SPONGE
BARNACLES
BIVALVES
PREDATORY SNAILS
STARFISH
FISH, BIRDS, SEALS
MICRO-FLORA
BROWSING MOLLUSCS
SEA ANENOMES
ATTACHED SEAWEEDS
SEA URCHINS
PRIMARY PRODUCERS
HERBIVORES
CARNIVORES

Figure 1-1

Principal Trophic Relationships in Rocky/Cobble Intertidal Habitat

2. EXPERIMENTAL DESIGN: ECOLOGICAL ASPECTS

2.1 INTRODUCTION

In this chapter we develop ecological design criteria for experimental oil studies in rocky intertidal habitats. Our goal is to establish a basis for choosing experimental hypotheses, treatments, units and measurements for performing ecologically and statistically valid field experiments to investigate effects of oil. What we call "design criteria" are intended to be useful guidelines for both field investigators and administrators in making decisions relating to experimental field studies of petroleum. Application of these criteria to specific design decisions depend on exact hypotheses to be tested (i.e., experimental objectives). We defer detailed examples illustrating such applications to Chapter 4.

Ecological design criteria are the basis for choosing experimental objectives, treatments, units and observations. They are largely qualitative and subjective and represent the knowledge and wisdom gained by experience working in the field. Our intent in this chapter is to distill from that experience, as reported in the literature, guidelines which can be useful in designing oil impact experiments. Our guidelines or criteria represent common themes which emerge for us in our review of the literature with the specific concern of designing oil impact field experiments. Our suggestions are not hard and fast rules, but rather guideposts. Each individual investigator's experience and judgment are always essential ingredients and we make no attempt to replace these.

In the following discussion we do not present a comprehensive literature review. Rather, we draw on existing reports to substantiate and document particular points of discussion. We do indicate those references that provide more comprehensive coverage of the literature. We begin with consideration of reported experiments in rocky shore habitats. With this backdrop we then turn to the particular problems of oil experiments.

2.2 EXPERIENCE WITH ROCKY SHORE EXPERIMENTS

In a survey of field experiments in intertidal habitats, Connell (1973) defines the ideal experiment as one in which the investigator manipulates one particular factor, while all others vary naturally. Although the investigator may only manipulate one variable, other causal factors may be structured into the experiment as treatments by the use of blocking and other devices. In theory all other naturally varying factors must be randomized out to prevent systematic bias. This is a difficult task even in small scale field systems. As Connell (1973) points out, it is difficult to manipulate a rocky intertidal system on a large scale, and so usually only small subsystems can be studied. The lack of control over naturually varying factors often prevents the establishment of true experimental controls or replicates. At best, the investigator may be able to observe untreated reference areas.

Connell (1973) reviews reported experimental manipulations of the physical environment and the biological environment. He concludes that these experiments demonstrate that the physical environment directly determines the distribution of marine organisms mainly when it reaches extreme values. In less extreme circumstances biological interactions are the principal factors. Connell's review provides substantial evidence for the feasibility and utility of performing field experiments which are ecologically and statistically valid. He also suggests several guidelines for improving field experiment methods. These include:

1. minimize experimental disturbances to the natural community;
2. establish "control" areas at the same time and closely adjacent to the experimental ones;
3. observe replicate control and experimental units.

Subsequent to the literature reviewed by Connell, we find recent articles by Paine (1974) and Dayton (1971, 1975) further illustrate the potential value of valid field experiments in rocky shore habitats. Both of these investigators focus primarily on the role of biological interactions in determining observed patterns of species abundance and distribution on rocky shores. The experimental methodology of these studies is important.

"A variety of experimental manipulations reveal an underlying dynamic nature" (Paine, 1974). These manipulations are motivated by hypotheses regarding well-defined functional subsystems of the whole community. Particular attention is devoted to guilds exploiting the common resource "primary" space, and the processes which allocate this resource. It is interesting and important to note that the strength of these experiments is largely due to the skillful use of subjective judgment in choosing treatments, observations, and experimental units. Sophisticated statistical techniques were not employed to choose the experimental design. In general, statistical design techniques assist in evluating cost and information content of alternative designs, by providing methods for assigning treatments to units and choosing the number of units to use. However, statistical methods cannot compensate for inadequate or improper selection of treatments, observations, or units.

We draw four conclusions about field experiments in rocky shores from our review of the studies discussed above. First, *it is feasible to conduct valid experiments in rocky shore habitats to test hypotheses about causal relationships governing observed patterns of species abundance and distribution.* A great deal of work is required in planning and executing the experiments, but the complexities and vagaries of field situations can be dealt with. A common characteristic of successful experiments we have reviewed is the focus of attention on particular physical and/or biological factors acting on particular species or groups of species (guilds) and the manipulation of specific relationships among these factors. For example, Paine (1974) selected the predator-prey relationship between the mussel Mytilus californianus and the starfish Pisaster ochraceus as a basis for experimental investigation. Experimental manipulation of this functional relationship reveals a predictable dynamic structure within the community. This emerging "model" of the rocky shore contrasts (Paine, 1974) with an alternate view that this community is relatively static and unpredictable, controlled primarily by external physical factors. This latter "model" is principally derived from descriptive surveys of rocky shore environments. In a comprehensive review, Connell (1972) demonstrates further the importance of direct manipulation of relationships in determining relative importance of causal factors governing patterns of species abundance and distribution in rocky shores.

Secondly, *there have been relatively few controlled manipulation experiments carried out and reported in the literature.* Corresponding to the lack of field investigations, is the apparently small number of investigators skilled in carrying out these experiments. The literature in this area is dominated by the names of Connell, Paine and Dayton. R. T. Paine's group at University of Washington appears to be the principal training ground for investigators skillful in these kinds of experiments.

Thirdly, *a critical problem in conducting these experiments is the time frame over which governing phenomena exhibit themselves.* Some of the phenomena of interest especially related to recovery processes occur on time scales of years, so that the rate of new information produced may be slow relative to current rates of resource development decisions. Irregularity of larval settlement and recruitment can be a particularly troublesome problem. An essential part of understanding impacts of oil spills is the process of community recovery. This is in part governed by recuitment success which can be highly variable from species-to-species and from year-to-year for a given species. Dayton (1971) concludes that his inability to confirm previous experimental results of Pisaster removal reported by Paine (1966) was the result of differences in Mytilus settlement during the years in which the experiments were conducted. Experiments designed to investigate these phenomena are likely to require several years of observations to yield ecologically and statistically valid results. There is little, if nothing, that can be done about this characteristic of field studies; it is an inherent property of the biological processes controlling natural community structure and function. Policy makers must learn to accept this reality.

Fourthly, *much remains to be learned about the processes governing species abundance and distribution in natural rocky shores, much less understanding the effects of oil.* The studies discussed above have provided important insights to the dynamic processes occurring in these systems, but each experiment must focus on a limited aspect of the whole community. Many phenomena remain unstudied. This lack of basic understanding severely limits the scope of hypotheses and experimental designs about oil which we can investigate in biologically and statistically meaningful ways.

2.3 REPORTED EFFECTS OF OIL: A BASIS FOR CHOOSING TESTABLE HYPOTHESES

An extensive volume of literature exists on the biological effects of oil. Many reviews exist. Interested readers are referred to NAS (1975), Boesch et al (1974), Moore and Dwyer (1974) and most recently Malins (1977) and Wolfe (1977). This latter volume provides an excellent review based on more recent results, which benefit from significantly improved experimental procedures and analytical methods, especially chemical. Much of the work reported prior to 1973 (marked by the NAS Workshop, see NAS, 1975) does not benefit from accurate, quantitative chemical analyses and therefore is of considerably less value. In spite of this extensive literature, much remains to be learned about the effects of oil. An assessment of research needs is presented in Wolfe (1977). However, sufficient experience exists to categorize the effects of oil and suggest types of hypotheses suitable for further experimental investigation.

It is important to recognize that oil and organisms are interrelated. Organisms modify the characteristics of oil and, of course, oil has effects on the plants and animals themselves. Biological transfers and transformations may profoundly alter both the chemical nature of petroleum hydrocarbons, and the distribution of hydrocarbons in the environment. Effects of oil are exhibited at many levels of biological organization: cellular up to community. In this report we distinguish among sub-organism, organism and population (community) level effects. The specific actions of petroleum hydrocarbons occur at the sub-organism level and are exhibited as changes in organism function, especially growth, reproduction and mortality. These effects propagate through populations and communities of organisms, modulated by the complex processes governing ecosystem structure. The initial response to oil exposure and subsequent recovery of species abundance and distribution are of primary interest in field studies.

Corresponding to the foregoing differentiations we use the following categories to describe the fate and effects of petroleum hydrocarbons:

1. Biotransformations
2. Lethal and sublethal organism level effects
3. Population/community effects

Biotransformations are the organism effects on the structure of petroleum hydrocarbons. By a variety of mechanisms organisms can take-up these chemicals. Metabolic processes may cause changes in the chemical structure of the hydrocarbons. Some organisms demonstrate significant storage of hydrocarbons and metabolic conversion products. The contaminating substances are discharged in different forms. In Wolfe (1977) several authors discuss details of biotransformation processes and review existing literature. Note that foodweb transfers and tissue contamination are observable effects of these biotransformation processes.

Exposure to hydrocarbons can have a variety of lethal and sublethal effects exhibited on whole organisms. Direct lethal effects can result from poisoning (toxicity) or mechanical smothering and interference with organism function. Water soluble aromatic hydrocarbons are the apparent cause of toxic effects (for example, see Rice, et al in Wolfe, 1977). Smothering is primarily caused by the higher molecular weight hydrocarbons, such as "tars." Sensitivity of individual organisms to petroleum hydrocarbons is a complex function of many factors including composition and amount of hydrocarbons, length of exposure and life stage. Accurately measuring toxicity of hydrocarbons is a difficult task in the laboratory (see, for example, Rice, Short and Karinen or Anderson in Wolfe, 1977). Predictions of organism responses in field situations are extremely uncertain. Little is known about the correlation between LC_{50} levels as measured in bioassay tests, and levels of hydrocarbons causing observable effects in the field.

Sublethal effects are abnormal responses of an organism to concentrations of hydrocarbons that are below lethal concentrations measured as the amount required to kill 50% of the test organisms in 96 hours (96 hour LC_{50} or 96 hour TLM). Anderson (in Wolfe, 1977) categorizes studies of sublethal effects as histological studies, physiological studies, behavioral studies and studies of growth and reproduction. Under physiological studies he includes investigations of respiration, osmoregulation, feeding and nutrition. We adopt the point of view expressed by Anderson (in Wolfe, 1977) that growth and reproduction are the most valuable parameters for measuring sublethal effects of oil. Sublethal effects not ultimately expressed as changes in organism growth and reproduction may be scientifically interesting, but are not likely to be environmentally significant, with one exception. Hydrocarbon contamination of tissues ("tainting") is of interest, no matter what other effects are exhibited.

Population and community effects are changes in patterns of species abundance, composition and distribution. They result from the propagation of lethal and sublethal effects on organisms through the ecological system. Community changes can also occur due to habitat alterations, such as rock substrate coating by oil, which prevents settlement and colonization of space by organisms. It is useful to recall that observed patterns of species abundance and distribution arise from complex processes involving physical and biological factors. Oil is an additional factor which is superimposed on naturally occurring processes. To fully understand the effects of oil, we must have some understanding of the cause-effect relationships which function in the absence of oil.

As a guide for further definition of testable hypotheses about effects of oil, we turn to results of field studies made following accidental oil spills. These observations suggest correlations between exposure to oil and changes in species abundance and distribution. These correlations can, in turn, lead to hypotheses; and experiments can be designed to test these hypotheses.

The most frequently reported observation of changes in species distribution patterns following oil spills is an alteration in abundance and distribution of algal species and related dominant herbivores. Nelson-Smith (1972) summarizes observation of algal-herbivore responses following the Torrey Canyon and Tampico Maru spills, several spills in Milford Haven, and field experiments conducted at the Orielton Field Station in England. Clark et al (1975) describe adverse effects of diesel oil spilled from the General M.C. Meigs on intertidal plants and on sea urchins. Thomas (in Wolfe, 1977) documents a long-term effect of bunker C oil spilled in Chedabucto Bay on fucoid algae. These effects have been attributed in different cases to toxic effects and non-toxic (smothering) lethal effects. In any case, recovery has typically involved a sequence of species recolonizing the affected areas. As described by Boesch et al (1974):

> *"Particularly if herbivorous invertebrates are killed, green algae (e.g., Enteromorpha and Ulva) will rapidly recolonize denuded intertidal rocks, producing a characteristic "green phase." This will gradually be replaced with a cover of brown fucoid algae, which are more robust. If the fucoids are not grazed they may form a thick cover that inhibits the reestablishment of the usual fauna of mussels and barnacles. The recovery of herbivore populations and consequent reestablishment of an algae/grazer balance are necessary for the complete recovery of the system." (p. 17)*

Wide spread mortality over many invertebrate taxa has only been reported in a few cases. Most notable are the Tampico Maru and West Falmouth spills of fuel oils. Also, high rates of mortality were reported from the application of toxic dispersants used in the clean-up of the Torrey Canyon spill. Effects of crude oil spills apparently are primarily the result of physical effects such as coating and smothering. Organisms show varying sensitivities to these mechanical effects. For example, Straughan (1971) reports that Chthamalus, a high intertidal barnacle species was completely covered by oil after the Santa Barbara spill and suffered high mortality. A taller barnacle species survived because it projected above the coating of oil. In field experiments with crude oil, Crapp (1971) concluded that experimental oiling of a rocky shore did not damage plants or animals, except where the thick "atmospheric residue" of crude oil physically dislodged periwinkles and other gastropods (e.g., Thais). Ranwell (1968) reports significant oil-smothering mortality among algae and invertebrate species along rocky shores following the Torrey Canyon spill.

Boesch et al (1974) suggests that organisms having a natural coating of mucilage or mucus, such as macroalgae and anemones, are less affected by oil because it does not adhere to their surfaces. North (1964) reports that a green anemone was one of the few surviving species following the Tampico Maru. At Santa Barbara certain algae were not as heavily damaged as intertidal surfgrass (Nicholson and Climbert, 1971). However, at Chedabucto, Thomas (in Wolfe, 1977) reports that oil was observed to adhere tenaciously to seaweeds from the upper intertidal zone, leading to extensive mortality to Fucus spiralis. Clark et al (1973) describe severe damage to the physical structure of a laminarian seaweed and tissue bleaching of several other intertidal algal species.

Observations of patterns and length of recovery from oil spills are more varied than reports of the extent of initial impacts. Recovery patterns depend on the extent of initial mortality, the persistence of the spilled oil, and community recolonization and development processes. In cases of widespread mortality such as the Tampico Maru, recovery can apparently take ten years or more. When less extensive mortality occurs, recovery depends on the particular species affected and their role in community structure. Clark, et al (1973) conclude from observations at Wreck Cove and at other spills that loss of algae can have profound effects on other organisms, not directly affected by the oil. Certain species of algae act as habitat formers, providing food and shelter for certain intertidal animals. Furthermore, disturbance of algal-herbivore balances can lead to wide-ranging shifts in species abundance and distribution. Unusually dense growth of algae occurred after the elimination of herbivorous gastropods following the Torrey Canyon spill (Smith, 1968) and the elimination of sea urchins following the Tampico Maru spill (North, 1964). Return to a distribution of species similar to the pre-spill conditions can take several years and depends on the success of larval settlement and re-establishment of adult populations of the herbivores.

The recovery process subsequent to the Chedabucto spill has followed a different pattern than other cases (Thomas, in Wolfe, 1977). Sporelings of fucoid algae have repeatedly settled in the upper intertidal zone but have not survived to an identifiable size. In Chedabucto Bay this zone is normally dominated by Fucus spiralis. Below this zone Fucus vesiculosus dominates. Following the spill the upper limit of this species was depressed and over a 4-6 year period has returned to successively higher levels, as surface oil cover decreased. Thomas (in Wolfe, 1977) attributes the delay in recovery of F. spiralis to persistent toxicity of oil. However, virtually no data is available on the chemical characteristics of the persistent hydrocarbons. Thomas (in Wolfe, 1977) also notes a "minor contraction" of the range of the periwinkle Littorina obtusata. This observation is attributed to the disappearance of its algal habitat in the upper part of its range.

We can find no reports of observation of recovery processes associated with the disturbance of herbivore-carnivore balances. Apparently no significant disruptions of carnivore population such as the starfish _Pisaster_ or the snail _Thais_ have been observed.

Persistent contamination of an area by oil may prolong recovery by causing a continuing mortality or by preventing recolonization. At Santa Barbara settlement of barnacle larvae were observed within three months of the spill. On the other hand, fucoid sporelings continue to die-off after settling in Chedabucto Bay seven years after the spill event. North (1964) reports that three months after the _Tampico_ spill the cove affected appeared "fresh and clean." However, chemical data are not available. Oil residues were observed for two and a half years following the wreck of the _Meigs_ (Clark et al, 1975).

Oil stranded on rocky shores may weather to an asphaltic residue on rock surfaces. Barnacles recolonized this surface at Santa Barbara (reported in Boesch et al (1974). However, Chan (1973) reports that recolonization did not occur on weathered oil. Blumer et al (1973) discuss the processes governing the weathering of Stranded Crude oil.

Very little data have been reported on the persistence of hydrocarbons in organism tissue following a single spill of crude oil. Clark et al (1975) describe long-term studies in a rocky shore habitat following a spill of fuel oil. They were unable to determine if persistent contamination for 2-1/2 years resulted from periodic recontamination or the original large petroleum uptake immediately following the spill from the _Meigs_. Several authors have reported on the persistence of oil contamination by #2 fuel oil at West Falmouth (see, for example, Michael, et al, 1975). Lee (in Wolfe, 1977) summarizes amounts of petroleum hydrocarbons in bivalves measured by several investigators. Based on these reports he suggests that bivalves can be useful for monitoring petroleum inputs. He observes that in an oil spill area or in areas of chronic oil pollution, the hydrocarbon composition of bivalves reflects the concentration and relative amounts of different hydrocarbons in water.

Gilfillan et al (1976) report laboratory and field measurements which suggest certain effects of hydrocarbon contamination of organisms on animal physiology. A carbon budget for soft-shell clams exposed for several years to #6 fuel oil was estimated and compared to an unoiled population. The net gain of carbon by the oiled population was 50% of that gained by the unoiled population.

The effects discussed above are not all suitable for study in field experiments. Large variability is characteristic of field situations and prevents the biological and chemical control necessary for elucidating cellular mechanisms such as associated with biotransformations of petroleum hydrocarbons. Monitoring hydrocarbon levels in tissues of organisms collected in the field is not unusual. However, cause-effect relationships at this level of biological organization are properly the subject of laboratory studies. As Connell (1973) argues:

> *"If questions are asked at the individual or lower levels,... then laboratory experiments are completely appropriate. It is only when the question concerns the behavior of the population in nature,...that laboratory experiments are less useful than field experiments." (p. 50)*

Following Connell's discussion *we believe that the primary focus of experimental studies of oil in the field should be on population and community effects.* Of course, such experiments may require measurement of organism and sub-organism variables, but *hypotheses to be tested should concern patterns of species abundance and distribution.* Understanding foodweb transfers will ultimately require field studies, but in our opinion, the current literature makes clear that too little is understood at sub-organism level to expect field experiments to yield significant information about causal relationships governing rates and routes of hydrocarbon transfer. However, field observation may provide data which are useful in choosing and designing appropriate laboratory experiments.

We find it useful to differentiate between two kinds of hypotheses about effects of oil on patterns of species abundance and distribution. These are hypotheses about: 1) "immediate" or "initial" impacts of oil; and 2) patterns of recovery following an oil spill.

Specific hypotheses regarding the initial impacts of oil are included within the following generic question.

- *What amounts and compositions of oil cause a detectable change in density of selected species?*

  This question could be restated as a hypothesis in terms of the reduction in density associated with different levels of oil. We use the word "oil" to refer to any hydrocarbon mixture. By applying "oils" of different hydrocarbon composition, toxic and non-toxic lethal effects can be differentiated. It is useful to select the type and levels of treatment in accordance with laboratory data. If LC_{50} data are available for the species under study, then treatments can include levels of oil corresponding to the known LC_{50}.

  For mobile species, reductions in density may occur from direct lethal effects or due to behavioral response to avoid the contaminated area. Differentiating such responses may require installation of experimental devices and/or specialized measurements to prevent or detect avoidance reactions.

Recovery processes encompass a much larger range of phenomena and are generally exhibited over time periods measured in years. We include questions regarding effects of long term chronic exposures under studies of recovery processes, because the length of time involved necessarily means that many of the same ecological phenomena are involved. Unlike the problem of initial impacts, no single generic question can be posed which can be used as a guide to devising specific hypotheses about recovery.

However, we can disaggregate the problem into several parts and identify questions applicable to particular species of interest.

- *Does the presence of oil affect the recruitment and settlement of individuals into contaminated areas? How long does full recovery take?*

  Note that if oil persists in the area at levels lethal to the species of interest, then the recovery process is delayed, because new recruits cannot survive to become mature adults. Also the degree of mortality suffered may influence the recovery process.

- *What is the response of populations not directly affected by the oil, but ecologically linked to species which are affected?*

For example, algal "blooms" have been frequently reported following reductions by oil in herbivore densities. Many aspects of the phenomena require further investigation, e.g., What is the response to various changes in herbivore density? Also, little is known about the response for other guild linkages.

- *How long do various types of oil persist as contaminants to various species?*

  Persistence of oil can be expected to alter recovery processes as mentioned above. Experiments may indicate determination of which species are affected chemically by persistence and which are affected physically by persistence.

- *How does the spatial scale of a disturbance affect the rate and sequence of recovery?*

  The spatial extent of an oil treatment may affect the process of recuitment. If recruits are derived from the local population, then the number of survivors may influence the rate of recolonization.

The foregoing questions represent generic aspects of the effects of oil spills which may be studied by field experiments. We consider them to be a set of general experimental objectives, for which specific designs may be developed. Examples of such specific designs are presented in Chapter 4.

2.4 OIL AS AN EXPERIMENTAL FACTOR

Identifying a testable hypothesis is only one step towards designing an experiment. We must also choose experimental treatments and units which yield suitable data for actual analysis and hypothesis testing. Several characteristics of oil complicate the field experiment problem and lead to difficulties in obtaining desired experimental treatments.

First, oil as it occurs in an oil spill is not a naturally varying factor. It is in this amount and composition a foreign substance in most rocky shores. Unlike manipulation of, say, predator density, an oil spill is a disturbance for which the community has no historical precedent. Although hydrocarbons are ubiquitous in the marine environment, organisms have not evolved in the presence of this factor in the amounts and compositions associated with spills. The fact that oil is a foreign substance does not invalidate properly designed experiments. However, choice of the experimental treatments, observations and units and interpretation of results may be more difficult because we have relatively little definitive knowledge with which to develop testable hypotheses.

Secondly, oil is pervasive, not selective in its effects. Oil directly and simultaneously may affect physical variables, (e.g., substrate area), the chemical environment, and the biological environment (e.g., increase/ mortality). Because of the pervasiveness of oil effects, designing experiments which allow particular relationships to be investigated appears to be extremely difficult. Oil cannot be indiscriminately applied and expected to exhibit only selective effects of interest to the investigator. Even when experimental objectives are carefully selected, it may not be possible to devise treatments and units which confine oil physically and chemically to desired spatial units or to only species within a given unit.

Another difficult characteristic of oil is its complex chemical composition. Oil is not a single uniform substance. It is a mixture of many different hydrocarbon compounds of different molecular weight and structure. As a result of different chemical properties (solubility, vapor pressure, etc.), each compound has different biological effects. Oils from different origins are different mixtures of hydrocarbons so their biological effects also vary. In order to assess the biological effects of oil and compare different oils, detailed chemical analyses must be obtained. The chemical problems and measurements associated with studies of biological effects of oil have been recently discussed by Rice, Short, and Karinen (in Wolfe, 1977) and other authors in the same volume. In short, *any experiment using oil as a treatment must include a measurement of hydrocarbon composition by type (e.g., paraffin, aromatic, etc.) and molecular weight throughout the duration of the experiment.*

Part of the chemical complexity of oil is its instability in natural environments. The volatility, solubility and biodegradability of lower molecular weight hydrocarbons leads to rapid "weathering" of spilled oil, i.e., rapid changes in the hydrocarbon composition of the spilled material.

As a result, an oil "treatment" is not constant; it is time-varying, further complicating the experimental problem. Again, carefully detailed chemical analyses are a necessary part of any investigation of the biological effects of oil. Chemical data throughout the course of an experiment provide the only way of accurately describing the kind and amounts of chemicals to which organisms are exposed.

Finally, we return to the problem of having relatively poor *a priori* information about causal relationships in rocky shore communities--with or without oil. As a result, our hypotheses are uncertain and the models upon which we base experimental designs are uncertain. Part of the design problem is to evaluate the trade-off between the likelihood of obtaining ecologically and statistically significant information and the potential value of the information, given the costs. If our *a priori* understanding of the phenomena is highly uncertain, the probability of obtaining significant results may be low for any acceptable experimental effort (cost). In this case we are forced into a slow, iterative process of learning in which many efforts may be unsuccessful and designs are largely *ad hoc*. A high rate of information return cannot be expected.

Based on the foregoing discussion, we are pessimistic about the potential for obtaining significant, valid results from field experiments using oil as a treatment. One approach is to design experiments to study oil effects, which do not actually include oil as an experimental treatment. The validity of this approach, which we explore in an example in Section 4.4.3, depends on the particular hypotheses under investigation. Another possible way to deal with the problem may be to carry out experiments, very limited in scope, in which the subsystems under study are restricted to experimental units which are of the order of square centimeters in area and encompass only a few organisms. By reducing the scale of the experiment, this approach may increase the controllability and reduce the variability of all factors in the experiment, including oil. Of course, inferences from such experiments are correspondingly restricted to more limited populations, further reducing the rate of information return, but increasing the validity of the observations. Determining spatial scale of oil treatment experiments is a necessary prerequisite to conducting impact and recovery experiments. The design of feasibility experiments is discussed in Section 4.4.2.

2.5 CHOOSING EXPERIMENTAL UNITS AND OBSERVATIONS

Given our experimental objective of investigating causal relationships governing patterns of species abundance and distribution in the presence of oil disturbances, we adopt an approach similar to Dayton (1975) and Paine (1974) for choosing experimental units and observations. We begin by dividing the community of rocky shore organisms into species groups ("guilds") according to their functional role in the community. Examples of guilds are the barnacle-mussel-anemone-sponge complex, large benthic algae, herbivores, detrital feeders and predators (starfish, Thais, etc.). The actual species constituting a particular guild may vary from location to location, depending on specific local conditions such as wave energy, slope, orientation and splash. Preliminary experiments at a site may be required to describe the species and their relationships in the absence of oil. Examples of experiments based on guild linkages are Paine's studies of the relationship between the mussel Mytilus Californianus and its predator, the starfish Pisaster ochraceus (Paine, 1974); and Dayton's studies of algal-herbivore-carnivore linkages in the intertidal zone (Dayton, 1971 and 1975). Focusing our attention on small groups of functionally related species is one way of limiting the scope of experiments and working with a defined subsystem of the rocky shore.

After choosing a subsystem for experimental manipulation, we must define the actual experimental units which will receive various treatments. Dayton (1971), Connell (1961) and Paine (1971) describe experimental devices such as wire cages and "roofs" for including and/or excluding certain species. These devices constitute experimental units and typically have an area of 1/64 to 1/4 m^2. Experimental units otherwise consist of standard square sampling quadrats equal in areal size to the manipulation devices. We have previously pointed out difficulties which may arise with confinement of oil to experimental units. We discuss this problem further in the context of example designs in Section 4.4.2.

The choice of quadrat or unit size is open to some question. Elliott (1971) points out that the choice of quadrat size is a compromise between statistical and practical requirements. He argues that, in general, smaller sampling units lead to more accurate and representative results. However, this depends on the pattern of spatial distribution among organisms. If organisms are randomly distributed in space, then all quadrat sizes are equally efficient in the estimation of population parameters (e.g., mean and variance). If clumping occurs, then the effect of quadrat size depends on the scale of clumping. The choice of quadrat size will depend on the scale of clumping one wants to observe. Typically quadrat sizes of 1/64 to 1/4 meters squared (m^2) have been found effective in rocky intertidal sampling. The larger size requires significantly more effort in sorting and enumerating samples in the laboratory than smaller quadrats of say 1/64 m^2. In the final analysis, the choice of quadrat size and shape will largely depend on the experience and judgment of the investigator, as applied to the particular vagaries of the site in question.

Given the kinds of hypotheses with which we are concerned the primary observations of interest are abundances of selected species within each experimental unit. In general, abundance may be measured by percent coverage, counts and/or wet weight.

The three principle methods of measurement are non-destructive measurement, photographic, and destructive collections. The occupation of space can be measured photographically, as described by Connell (1970) and Dayton (1971). Because photographs provide a permanent record of numbers, sizes, and surface area covered by many species, we recommend their use. Destructive collections should be avoided because of the gradual elimination of the experimental units and possible effect of the measurement method itself on the community structure.

Percent coverage measurements, which are significantly easier than direct counts, may be adequate for numerically dominant sessile species. This is particularly true for dense stands of small organisms, because percent coverage measurements are less variable than numbers of individuals in such cases. For example, mussels may cover virtually 100% of the space in a particular area and have densities as high as several thousand individuals per meter. However, counts of individuals occurring within a quadrat within the zone can show high variability, because individual organisms can be quite small (<1.5 cm) and utilize "secondary" space, as well as "primary" space. Data reported by Paine (1974) and Zimmerman and Merrill (1976) demonstrate the extent of variation characteristic of mussels.

In spite of these advantages, percent coverage measurements are not always appropriate. Rare species or small organisms which do not occur in dense clusters may be missed using percent coverage. Where such species are of interest, enumeration by counting is necessary.

In addition to measurements of species, observations of other variables may be necessary. In Section 2.4 we noted the importance of measuring amount and composition of hydrocarbons in substrates, water column and organism tissues. Other environmental variables such as tidal elevation exposure, temperature, etc. may also be of interest in particular experiments.

2.6 SUMMARY OF ECOLOGICAL ASPECTS OF DESIGN

In this chapter we have identified criteria for choosing experimental objectives, treatments, measurements and units for conducting field experiments on the effects of oil. These choices constitute what we call the ecological design guidelines and depend on the objective of the study, the intended users of the data and the investigator's understanding of the specific site of the experiment. Because we have not narrowed our analysis to a specific experimental site and because of the multitude of possible specific hypotheses to be tested we have not attempted to exhaustively enumerate possible designs. Rather, our intention has been to provide some guidelines for designing a particular experiment. The statistical design problem, which is in part dependent on these choices, is described in Chapter 3; and specific examples are given in Chapter 4. The following statements summarize the design guidelines we have explored in this chapter.

- *Choose a small sub-system of the rocky shore habitat which is observable and controllable.*

  Sub-system size is measurable taxonomically (number of species included) and spatially. Experiments should be conducted on a scale such that treatments and manipulations are under the investigators control. If experimental, untreated "controls" cannot be established, the experiment should not be undertaken. The number of species included in the experiment should be kept small as possible. This is a highly subjective and difficult decision (Paine 1974). Focusing interest on species guilds can facilitate this decision.

  The major disadvantage of reducing the size of the experimental system is that generalizations to larger systems are necessarily restricted. Therefore, the rate of information return from the experiments is reduced. As understanding of the processes governing the system increases, larger subsystems can be defined for experimentation.

- *Minimize the experimental disturbance to the natural system.*

  Connell (1973) discusses the importance of this criterion for field studies. Unnecessary disturbances can obscure effects of interest and confound experimental results. We have previously discussed problems associated with using oil as an experimental treatment. The difficulty is in part due to the indiscriminate nature of the disturbance caused by oil.

- *Focus experimental objectives on population/community effects, not at the level of the individual or lower.*

  We have discussed this criterion previously in Section 2.3. We reiterate that measurement of organism and suborganism variables may be required, but that hypotheses to be tested should concern patterns of species abundance and distribution.

- *Experiments which include oil treatments should include careful measurement of hydrocarbon composition and levels in substrates, water column and organism tissues.*

  The complex chemical properties of oil can only be accounted for by careful monitoring throughout an experiment. In fact, the continually changing chemical characteristics of oil make it difficult to control. It is not a constant experimental treatment. Observed biological responses must be related to the organism's *exposure,* not the original treatment of oil.

- *Formulate testable hypotheses about specific species, relationships and treatments.*

This may be the most important guideline in this list. Refining the experimental objectives into statistically testable hypotheses will naturally elucidate the treatments and measurements required in the experiment. Therefore, the feasibility of the experiment can be more readily evaluated.

3. EXPERIMENTAL DESIGN: STATISTICAL ASPECTS

3.1 INTRODUCTION

In principle, it would seem that most of the statistical concepts of traditional experimental design theory could be easily applied to intertidal ecological investigation. Systematic methods for laying out sampling networks and collecting experimental data have been used for years in agricultural research studies which are superficially similar to intertidal field experiments (Fisher, 1926; Fisher, 1935). Natural marine ecosystems are, however, much more variable and difficult to control than the cultivated ecosystems of agricultural experimentation. This situation is complicated further when petroleum is introduced, either intentionally or accidentally, into the field experiment. It is practically impossible to control or confine petroleum treatments to the simple sampling grids and quadrats used in agricultural studies. It is difficult to even define the extent of spilled oil as it spreads across the experimental site and undergoes physico-chemical changes. These difficulties can limit the applicability of traditional sampling methods to petroleum impact studies, requiring the statistically-minded experimenter to be particularly creative in his approach to experimental design.

Because of the potential design problems associated with petroleum-related experiments, it is appropriate to pose a few pointed questions about the relevance of statistical design criteria:

- To what extent can traditional experimental design methods be applied to intertidal and particularly to petroleum-related field studies?
- What standards must be met for a particular intertidal experiment to be statistically valid?
- How much effort (or cost) must an investigator spend to verify that a particular ecological impact is statistically significant?

Before these questions can be properly answered (they are reviewed in Section 3.4), we must examine more closely the techniques offered by experimental design theory and the unique problems associated with intertidal experimentation.

It is best to distinguish at the outset the statistical term *controlled experiment* from other terms commonly used in ecological field research. To the statistician, a controlled experiment is a study designed to identify *causal* relationships between observed dependent variables (e.g., species counts) and various independent variables (e.g., tidal elevation, time of year, etc.) thought to affect these observations. Although causality can rarely be established beyond a doubt, careful experimental design does permit an investigator to *infer* it with reasonable likelihood. The requirements of experimental statistics usually imply that some or all of the independent variables investigated in an experiment be manipulated (i.e., controlled) so that their influence may be clearly defined. This aspect of experimentation is closely related to the functionally-oriented approach to ecological research outlined in Chapter 2.

The frequent emphasis on controlled experimentation found in the statistical literature is not intended to minimize the importance of statistical considerations in other type of field studies. *Sample surveys* use statistical guidelines to derive sampling layouts for estimating the abundance and distribution of an existing population, without regard for causal factors (Kendall and Stuart, 1976, Chap. 39). *Analytical surveys* (a term used by Cochran, 1953, Chap. 5) rely on statistical concepts to provide layouts which can detect significant differences between two or more populations. Both types of surveys are important in certain phases of petroleum impact research and both are considered in more detail later in this chapter. We do, however, place more emphasis on the design of controlled experiments because we believe such experiments must eventually be carried out before the effects of oil in the natural environment can be reliably predicted. The methods and feasibility of survey sampling are well understood in benthic environments (e.g., Elliott, 1971) but controlled experimentation in this environment has been less extensive (Dayton, 1971; Connell, 1973). The concepts of control and experimental manipulation clearly need to be more carefully examined before actual petroleum perturbation experiments are attempted.

Both *experimental design* and *survey design* are concerned with the selection of sampling procedures and networks which maximize the amount of information obtained in the field for a given expenditure of effort. The particular type of information desired differs somewhat from case to case, but the basic principle of *optimizing* information return is quite general. For our purposes, an optimal design is one which gives better information than other available designs having comparable cost. The quality of experimental information is usually, but not exclusively, measured in terms of statistical indicators such as the *risk* associated with a decision or the *uncertainty* associated with an estimate. In abstract terms, *the problem of finding an optimal experimental design can be reduced to a search for controls, sampling layouts, replication strategies and experimental protocols which minimize statistical risk or uncertainty.*

This rather mathematical perspective should not overshadow the need for good biological intuition and everyday common sense in experimental design. The statistician's evaluation of risk and uncertainty is no better than the mathematical framework he uses to describe the experiment. If mathematical assumptions made for convenience or for lack of conflicting information are inappropriate, then some important practical design considerations may be overlooked. This is particularly true in the intertidal zone where the introduction of petroleum may have unanticipated impacts. As seen in our subsequent discussion there is no statistical merit in following a standardized experimental design in the interest of objectivity when casual observation reveals that a more innovative sampling procedure can yield better information. This point is repeated frequently throughout the report.

Given that an investigator wishes to obtain the most information possible for a given reserve of experimental resources (i.e., money, time, manpower, etc.) it is natural to ask whether he should allocate all his resources to a single comprehensive experiment or distribute them over a series of smaller related experiments. The latter approach, known as *sequential experimentation*, provides considerably more flexibility as well as allowing the researcher to refine his experimental designs as new field data gradually become available.

The best way to select a series of sequential experiments depends not only on the investigator's objectives but also on the type of ecosystem and perturbation under study. One possible approach to sequential design is outlined in the Zaikof Bay example of Chapter 4. There the following discrete experiment phases are defined:

1. *Preliminary survey* of the experimental site to determine abundance distributions, community composition, and species diversity (this information may be available in the literature for some sites).
2. *Initial manipulation experiments* designed to identify in a qualitative way the factors responsible for observed natural patterns of abundance (in the absence of petroleum).
3. *Oil feasibility experiments or surveys* to establish the feasibility of applying and controlling petroleum as an experimental treatment.
4. *Oil impact experiments* designed to establish the extent to which oil is an influence on the temporal and spatial variability of selected species.
5. *Recovery experiments or surveys* for a long-term follow-up of the effects of an intentional or accidental spill.

This example is mentioned here only to illustrate how the sequential philosophy may be applied to a practical oil spill impact assessment program. Since there are few generalized guidelines for sequential design (except in situations much simpler than intertidal research), the selection of the experimental phases of a sequential program must rely heavily on the investigator's judgment. Once the broad outlines of a program are established, more precise statistical methods can, in principle, be used to work out the details of each component experiment or survey. This chapter reviews relevant statistical concepts to determine their practical applicability to intertidal field studies. Chapter 4 illustrates how these concepts can actually be applied to specific experiments or surveys included in a sequential research program.

3.2 FUNDAMENTAL CONCEPTS

3.2.1 Experimental Models

In order to apply statistical concepts to experimental design and analysis we must have a clear and precise description of the experiment which identifies the measurements and design parameters available to us as well as potential sources of uncertainty and error. Such a description is provided by a *mathematical model* which explicitly relates the observations of interest (dependent variables) to various environmental factors (independent variables) and error sources. One of the simplest examples of a practical experimental model is provided by the following equation, which decomposes a single observation (y_{ij}) into several parts:

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{3-1}$$

where y_{ij} = observation at quadrat (i,j)
μ = mean of all observations
α_i = effect of the i^{th} level of an environmental factor thought to influence y_{ij}
e_{ij} = residual error (portion of y_{ij} not accounted for by μ or α_i) for replicate j.

In an intertidal context, the observation y_{ij} could represent the abundance of a particular species (in number of individuals/m^2) in the j^{th} quadrat located at tidal elevation i. The single experimental factor of interest would then be tidal elevation and the effect of elevation at level i would be α_i. The term *level* here refers to a particular value or classification of the environmental factor of interest. Some typical examples are:

Level Index	Tidal Elevation (above MLLW)
i = 1	0 - 1 m.
i = 2	1 - 2 m.
i = 3	2 - 4 m.
i = 4	4 - 6 m.

Level Index	Limpet Manipulation
$i = 1$	limpets not excluded
$i = 2$	limpets excluded

It is evident from these examples that a factor can be classified in a fairly quantitative way (as with the tidal elevation increments listed) or in a simpler qualitative way (limpets excluded or not). The decomposition of observations into mean, effect and error provides us with a way to distinguish the influence of particular experimental manipulations (factor levels) from other extraneous variations in abundance. Although we do not know the true values of μ and α_i, we can estimate them from the observations y_{ij} (if we have enough observations). These *effects estimates* can then be used to confirm or refute hypotheses about the relative importance of the experimental factors under investigation.

The terminology used in the previous paragraph (e.g., factors, levels and effects) is usually associated with controlled experiments and, in fact, a concern with causality and control is implicit in the entire discussion. As might be expected, a comparable set of terms is available for the model variables associated with survey studies. Since analogies between controlled experimentation and survey studies can be informative, it is useful to repeat Equation (3-1) with a survey-oriented set of definitions:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where y_{ij} = a sample of the population of interest drawn from quadrat (i,j)

μ = mean value of the population

α_i = departure from the mean in stratum i

e_{ij} = residual error for stratum i and replicate j

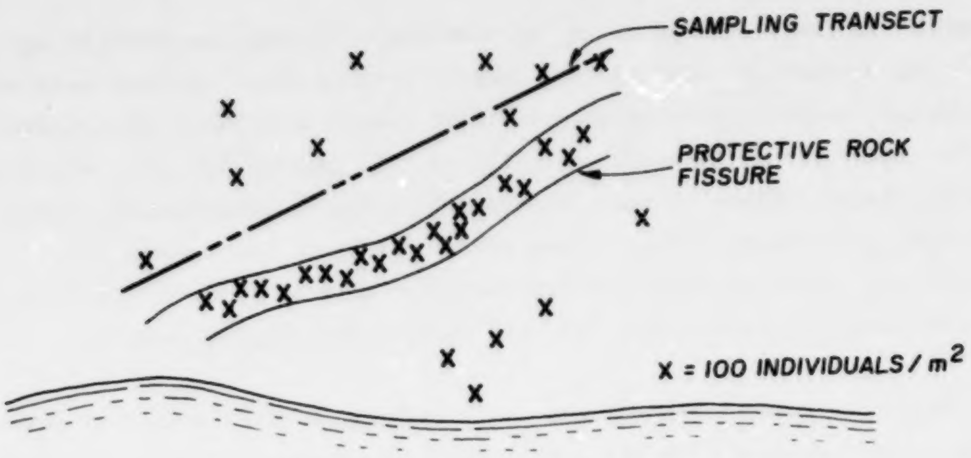
Here the term *stratum* refers to a sub-group of the larger population which is thought to be distinguished in some way. To clarify this, suppose we are interested in estimating the average abundance of a species (in number of individuals/m^2) at a particular site. If the species of interest tends to cluster in strata defined by wave exposure, topographic features, or other factors, it is likely that our abundance estimate will be highly dependent on the strata selected for sampling. If sampling quadrats are laid out along a transect parallel to a terrain feature which shelters much of the population, we will probably underestimate average abundance (see Figure 3.1a). If the transect runs directly along the terrain feature but does not cross less populated regions of the site, we will probably overestimate average abundance (see Figure 3-1b). These difficulties are best avoided if samples are collected from *both* transects (or strata) shown in the figure. The estimated population mean (μ) will then be more representative.

The above comments reveal that *both* controlled experiments and sample surveys are concerned with the estimation of unobservable parameters such as μ and α_i. The particular methods used for selecting sampling strategies and for computing estimates vary (see Kendall and Stuart, 1976, Chaps, 35-40 for an extensive discussion), but the basic concern with estimation is universal. Consequently, this chapter stresses general concepts of statistical estimation, decision making, and risk evaluation which can be applied equally well to controlled experimentation and survey sampling*.

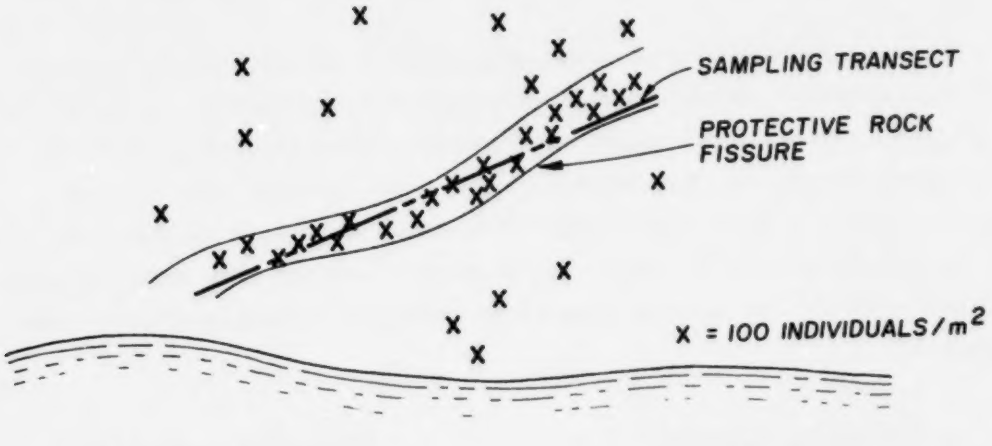
The simple single-factor model presented in Equation (3-1) includes only one dependent variable (observable) and one independent variable (factor). In general, a much larger number of dependent variables and factors may be of interest, requiring what is known as a *multi-variate, multi-factor* approach. This is best illustrated if we consider a list of potentially relevant variables for a typical field study concerned with the abundance of sessile species in a rocky intertidal community (example adapted from Chapter 4):

Dependent Variables	Independent Variables
1. *Balanus* abundance	1. tidal elevation
2. *Rhodymenia* abundance	2. *Katherina* abundance
3. *Odonthalia* abundance	3. carnivorous gastropod abundance
4. *Mytilus* abundance	4. limpet abundance

**Statisticians will note that our treatment of survey sampling assumes that the underlying population is infinite. This convenient assumption is probably justified in most intertidal studies since the total number of individuals present is much larger than the number sampled.*


SAMPLING TRANSECT
PROTECTIVE ROCK FISSURE
X = 100 INDIVIDUALS / m²

a) Underestimation of Average Abundance
(Population Mean)


SAMPLING TRANSECT
PROTECTIVE ROCK FISSURE
X = 100 INDIVIDUALS / m²

b) Overestimation of Average Abundance
(Population Mean)

FIGURE 3-1

An Example of Problems Encountered in
Non-Stratified Sample Surveys

Dependent Variables	Independent Variables
5. Collisella abundance	5. substrate type
6. Nucella abundance	6. time
	7. site location
	8. site exposure (exposed, sheltered
	9. topography (flat, moderate, steep)

The dependent variables listed on the left are the measures of community function of particular interest in this example. The independent variables listed on the right are environmental factors which may account for abundance variations or which may provide a good basis for sample stratification. In theory, *all* of the factors could affect all of the dependent variables in some way. Practical limitations of time, money and manpower require, however, that we focus on only a relatively small subset of the factors identified as being potentially important. This subset of factors is explicitly included in the experimental model.

Potentially important factors which are omitted from the model for reasons of convenience may still have a significant effect on monitored biological variables. An experiment which includes predator abundance in its model but ignores variations in site exposure may falsely attribute variations in prey abundance to predator pressure. In such cases, the ignored factor (site exposure) is called a *nuisance variable*. Methods for minimizing the detrimental effects of nuisance variables are discussed later in this chapter.

If several dependent variables and factors are included in the experimental model, a set of scalar equations similar to Equation (3-1) is needed to describe the experiment. This is easily seen if we first consider a univariate (single dependent variable), multi-factor model and then generalize to the multi-variate case. If time and site location are selected as the dominant factors in the list presented earlier, and if only one species (e.g., Balanus) is observed, then the following univariate, two factor (or "two-way") model provides a mathematically complete description:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (3\text{-}2)$$

where y_{ijk} = Balanus abundance in quadrat (i,j,k)
μ = observation mean
α_i = main effect due to the time factor (level i)
β_j = main effect due to the site location factor (level j)
γ_{ij} = effect due to interactions between time and location (levels i and j, respectively)
e_{ijk} = residual error for the k^{th} replicate of the (i,j) combination of factor levels

Each combination of factor levels (i,j) defined in the experiment constitutes a *factorial treatment*. One complete set of factorial treatments (all possible combinations of factor levels) forms an experimental replicate.

Cochran and Cox (1957) codify factorial treatments with a convenient notation which assigns an integer to each factor level. Any treatment may then be written as a sequence of digits such as:

A B C D E ← Factor
2 1 2 0 1 → Level

With this notation, 0 indicates the first factor level, 1 the second factor level, etc. In the above example we may define the complete set of factorial treatments with a simple two-dimensional table:

Table 3-1

Factorial Treatments for a 2 Factor Controlled Experiment

		Site Location Factor	
		Control Area (Level 0)	Petroleum Spill Area (Level 1)
Time Factor	Before Spill (Level 0)	00	01
	After Spill (Level 1)	10	11

Note that an experimental control (00) is naturally included in this problem formulation as one of a number of alternative treatment combinations. *Factorial experiments which include natural or background conditions as one of the levels in each classified factor will always have a control* represented by the 00...0 treatment.

With the notation described above we can readily write the model equations for a single replicate (k=1) of Balanus observations:

$$\begin{aligned}
y_{001} &= \mu + \alpha_0 + \beta_0 + \gamma_{00} + e_{001} \\
y_{101} &= \mu + \alpha_1 + \beta_0 + \gamma_{10} + e_{101} \\
y_{011} &= \mu + \alpha_0 + \beta_1 + \gamma_{01} + e_{011} \\
y_{111} &= \mu + \alpha_1 + \beta_1 + \gamma_{11} + e_{111}
\end{aligned} \qquad (3\text{-}3)$$

The size of any factorial experiment can be easily determined if the number of levels for each factor is noted. The factor example above can be described as a 2x2 experiment or, in even more concise notation, a 2^2 experiment. A four factor experiment having two, three, four, and four levels for each respective factor could be described by the notation $2\text{x}3\text{x}4^2$. This simple system can be extended to cover any factorial study.

In general, the total number of equations (i.e., samples) associated with a multivariate experiment is:

$$N = SK \prod_{i=1}^{F} I_i \qquad (3\text{-}4)$$

where I_i = number of levels for factor i

F = number of factors

K = number of replicates

S = number of dependent variables

Rather than write all of these equations individually, statisticians usually work either with a generalized scalar expression such as Equation (3-2) or with a matrix equation such as:

$$Y = X\Theta + e \qquad (3\text{-}5)$$

where Y = a matrix of observations
Θ = a matrix of treatment effects
e = a matrix of residual errors
X = an observation matrix which assigns treatment effects to observations

This rather abstract notation is easily visualized if the elements of each matrix are explicitly defined. To illustrate the way of dealing with multivariate experiments, suppose that the above 2x2 example is extended to include Rhodymenia as well as Balanus. The matrix expression is then:

$$\underset{Y}{\begin{bmatrix} y^B_{001} & y^R_{001} \\ y^B_{101} & y^R_{101} \\ y^B_{011} & y^R_{011} \\ y^B_{111} & y^R_{111} \end{bmatrix}} = \underset{X}{\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}} \underset{\Theta}{\begin{bmatrix} \mu^B & \mu^R \\ \alpha^B_0 & \alpha^R_0 \\ \alpha^B_1 & \alpha^R_1 \\ \beta^B_0 & \beta^R_0 \\ \beta^B_1 & \beta^R_1 \\ \gamma^B_{00} & \gamma^R_{00} \\ \gamma^B_{10} & \gamma^R_{10} \\ \gamma^B_{01} & \gamma^R_{01} \\ \gamma^B_{11} & \gamma^R_{11} \end{bmatrix}} + \underset{e}{\begin{bmatrix} \varepsilon^B_{001} & \varepsilon^R_{001} \\ \varepsilon^B_{101} & \varepsilon^R_{101} \\ \varepsilon^B_{011} & \varepsilon^R_{011} \\ \varepsilon^B_{111} & \varepsilon^R_{111} \end{bmatrix}} \qquad (3\text{-}6)$$

The superscript B's and R's refer, respectively, to Balanus and Rhodymenia. Also, the means for each species (μ^B and μ^R) are incorporated into the effects matrix Θ. The complexity of this expanded expression makes the convenience of compact matrix equations such as Equation (3-5) obvious. We will find matrix notation to be particularly useful when we review estimation concepts in the next section.

The possibility of using multi-variate, multi-factor experimental models allows us, in principle at least, to account for community effects and species interactions in an experimental design. In practice, community effects may not be well enough understood to justify complex and rather speculative models. As we shall see later in this chapter, factorial models should generally be kept simple so that each complete block or replicate of sampling quadrats can be confined to a reasonably uniform region. If interactions between even five or ten species are included in a community-oriented model, the number of factorial treatments and sampling quadrats required for each replicate can become very large (greater than one or two hundred). One way to deal with this problem is to develop aggregate community indices (such as diversity) as surrogate observables. This can reduce a multi-variate factorial problem to a uni-variate one, with a corresponding loss in the level of detailed information obtained. Another approach is to restrict the experiment to a few key species or species groups which have a particularly important functional role in the community. The relative merits of these methods for simplifying experimental designs depend largely on biological considerations which are beyond the scope of this chapter (they are briefly discussed in Chapter 2). From a statistical point of view, the important point is that the experimental model should be kept relatively simple and straightforward. Refinements can be made gradually as more is learned about the particular intertidal community under investigation.

The linear factorial model provided by Equation (3-5) is not by any means the most general mathematical model available for describing field experiments. It is the one most thoroughly covered in the statistical design literature as well as the one best suited to most intertidal impact studies. A few important generalizations of the model are worth mentioning, however, since they may be useful in particular applications.

The practice of classifying the factors of an experimental model into a small number of discrete qualitative levels (e.g., limpets present, limpets absent) is appropriate when more refined quantification is impractical. Occasionally, an experimental factor may be defined on a continuous measurement scale which offers more accurate quantification. In such cases the factor is called a *concomitant variable* and the discrete effects of the classified experimental model are replaced by regression coefficients. A commonly encountered example of this occurs when the initial abundance of a species (measured prior to any experimental manipulations) is used as a concomitant variable. If measurements of initial abundance are added to the single factor experiment described by Equation (3-1) the following model results:

$$y_{ij} = \mu + \alpha_i + ay^0_{ij} + e_{ij} \qquad (3\text{-}7)$$

where y^0_{ij} = initial species abundance
a = linear regression coefficient relating y^0_{ij} to y_{ij}

The unknowns in this model are μ, the treatment effect α_i, and the regression coefficient a. Both y_{ij} and y^0_{ij} are measured in the field.

Concomitant variables and regression terms may be incorporated into the general matrix model of Equation (3-5) if the regression coefficients are included in the matrix Θ and the concomitants are placed in the appropriate elements of the matrix X. The regression relationships used may include nonlinear functions of the concomitant variable. It is interesting to note that the use of time as a classified factor in the example of Equation (3-2) is an alternative way of including initial abundance in an experiment. There the measurements y_{001} and y_{011} at the control and impacted site are actually equivalent to concomitant initial abundance observations. The regression approach is usually better when there is good reason to believe that a simple linear or polynomial relationship between the dependent variable and noncomitant variable exists. The qualitative classified approach is preferred when no simple mathematical connection between concomitant and dependent variables seems appropriate. Often it is wise to try both approaches when experimental data are being analyzed.

Another generalization worth mentioning is the possibility of allowing the effects of a classified model to be random variables. Such *random effects* models are commonly used in the analysis of experimental results but are much less commonly used in experimental design. Some of the problems associated with random and mixed effect models are briefly discussed in Appendix A (we suggest that the reader refer to this Appendix only after reviewing Section 3.3). Unless otherwise noted, all the experimental models used in this report are classified, linear, fixed effect, factorial models.

Although most of the above discussion of factorial models has been from the perspective of controlled experimentation, the basic concepts are equally applicable to survey studies. Sample surveys which are stratified in more than one way (e.g., on tidal elevation and time) are comparable to multi-factor experiments, although they are not as commonly encountered in the statistical literature. Generally, the models and procedures used in biological surveys are simpler than those used in controlled experiments, although there are occasional exceptions. The interested reader is referred to Kendall and Stuart (1976, Chaps. 39 and 40) and to Cochran (1953).

3.2.2 Estimation of Treatment Effects and Population Parameters

The mathematical models described in the previous section provide a way to conceptually distinguish the effects of experimental manipulations (treatments) from extraneous environmental disturbances. Since treatment effects are themselves unobservable, their magnitudes and relative importance must be inferred from estimates based on available field measurements. The quality of these estimates depends both on the validity of the experimental model and on the accuracy of the field measurements.

Our simple representation of the experimental process combines model and measurement errors in one term, the residual error, which may be treated conceptually as if it were a completely random source of estimation uncertainty. Much of the "randomness" of this error source would disappear if we had a better understanding of the ecological processes at work at the experimental site. And, in practice, residual errors generally decrease as more realistic models are used to explain observation variability. But once the model and experimental layout are specified, the residual error may be treated as an unexplainable, totally random source of uncertainty.

Probability theory allows us to describe such random variables with probability densities such as the one sketched in Figure 3-2a. Here the bell-shaped curve specifies the likelihood (or probability) that a *particular* residual error (e_{ij}) will fall within a given range of values. The single-factor model equation:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

allows us to relate the probability density of y_{ij} (the observed dependent variable) directly to the probability density of e_{ij}. The y_{ij} density, shown in Figure 3-2b, is simply centered on the unknown value $\mu + \alpha_i$. Since y_{ij} is random by virtue of its dependence on e_{ij}, any estimate of μ or α_i based on y_{ij} will also be random. If, for example, μ is estimated by averaging all available measurements, then its estimation equation is:

$$\hat{\mu} = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij}$$

where

$\hat{\mu}$ = estimated value of μ

I = number of factor levels

J = number of replicates at each factor level.

The probability density associated with this randomly distributed estimate is shown in Figure 3-2c. It is narrower (has a smaller variance) than those of e_{ij} and y_{ij} because the mean estimate takes advantage of redundant measurement information. Generally, "good" estimation procedures yield estimate probability densities which are centered on the true value of the estimated parameter and which have the smallest possible variance. A detailed explanation of the basic principles of estimation theory may be found in Kendall and Stuart (1973, Chaps. 17-21).

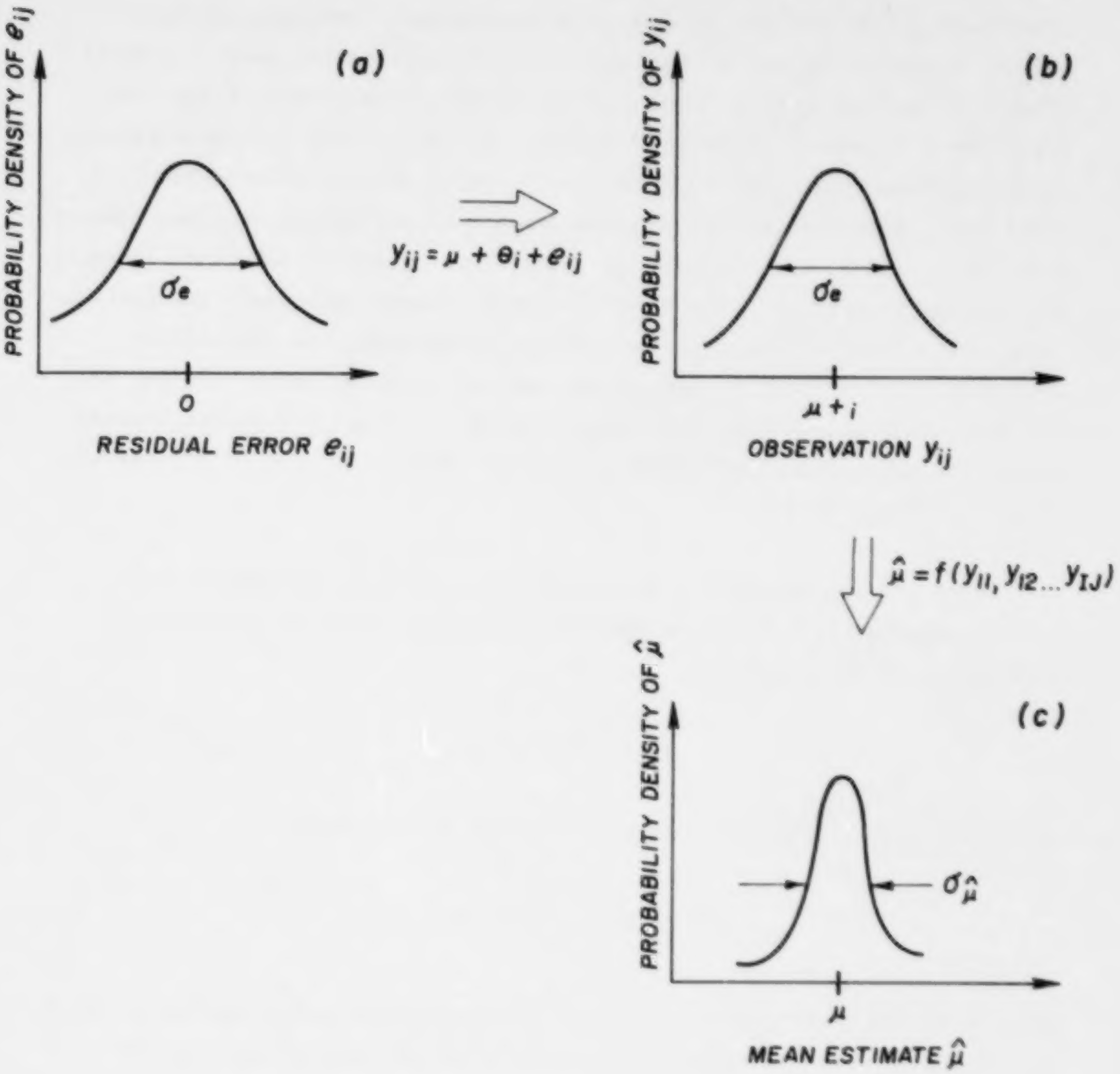

(a)
PROBABILITY DENSITY OF e_{ij}
σ_e
0
RESIDUAL ERROR e_{ij}
$y_{ij} = \mu + \theta_i + e_{ij}$
(b)
PROBABILITY DENSITY OF y_{ij}
σ_e
$\mu + i$
OBSERVATION y_{ij}
$\hat{\mu} = f(y_{11}, y_{12} \ldots y_{IJ})$
(c)
PROBABILITY DENSITY OF $\hat{\mu}$
$\sigma_{\hat{\mu}}$
μ
MEAN ESTIMATE $\hat{\mu}$

FIGURE 3-2

Relation of Residual Error, Observation,
and Effect Estimate Probability Density Functions
for a Fixed Effect Model

The estimation techniques most relevant to experimental design applications tend to fall into three categories--Bayesian estimators, maximum likelihood estimators and least squares estimators. Bayesian estimators provide a convenient way to deal with *a priori* information about treatment effects since they rely on a prior distribution of reasonable values for the effects matrix Θ. Maximum likelihood estimators have desirable minimum variance properties which apply when no *a priori* effects information is available. Both Bayesian and maximum likelihood estimators require, however, that the residual error probability density be completely specified (usually the density is assumed to be Normal). Least squares estimators are particularly attractive because they are density-independent, are intuitively appealing, and are usually easy to implement. Although each technique has its own strong and weak points, least squares is probably the best overall choice for most ecological research studies, largely because of its convenience and familiarity.

Least squares estimators are designed to minimize the discrepancy between observed and estimated treatment effects. When the linear model of Equation (3-5) is used:

$$Y = X\Theta + e$$

this discrepancy is measured by the weighted squared error:

$$S(\hat{\Theta}) = (Y - X\hat{\Theta})^T W(Y - X\hat{\Theta}) \tag{3-8}$$

where $\hat{\Theta}$ is the least squares estimate of Θ and W is a weighting matrix which may be adjusted to account for the relative uncertainty of the various measurements (Kendall and Stuart, 1973, discuss in Chapter 19 the relationship between W and the statistical properties of $\hat{\Theta}$).

As demonstrated in Equation (3-6), the residual error and other variables of the linear model are matrices when the model is multi-variate. In such cases, the weighted squared error is also a matrix. The elements of this matrix depend not only on the discrepancy between estimates of each dependent variable and observed values but also on complicated cross terms which are not related to obvious goodness-of-fit properties. Scalar quantities such as the matrix determinant provide a way to measure the magnitude of $S(\hat{\Theta})$ but they can be computationally troublesome. Considering the overall objectives of our investigation of intertidal experimental design, we feel it is best to restrict our attention to univariate statistical techniques which are much simpler to use. Readers interested in pursuing multivariate methods further are referred to Anderson (1958) and Rao (1952).

As noted in Section 3.2.1, a multivariate experimental design problem may be converted to a much simpler univariate problem if the various species abundances of interest are lumped into a single variable, such as a diversity index, which then becomes the dependent variable of a univariate multi-factor model. Alternatively, if the original species abundances are assumed to be independent, the original S-variate problem may be decoupled into S separate univariate problems, one for each species. Although the independence assumption may not always be biologically justifiable it does allow us to develop a practical approach to experimental design which is simple enough to be widely applied. Even if we adopted a coupled multivariate approach, we would still be forced to estimate unknown cross correlation coefficients, a process which could introduce errors comparable to those resulting from the independence assumption.

When the estimation problem is reformulated in univariate terms (either through decoupling or through the use of an aggregated index), the weighted squared error becomes an intuitively reasonable scalar index of estimator performance. In most applications, the least squares estimate Θ is chosen to minimize $S(\hat{\Theta})$ subject to certain constraints which reflect side conditions or hypothesized relationships among the model parameters which are imposed by the investigator. Often such constraints are needed in order to insure that the minimizing estimates are unique.

A simple example of this requirement is revealed by an inspection of the two factor model of Equation (3-2):

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

If each factor has two levels (2x2), the number of unknown parameters in this model is:

1	+	2	+	2	+	4	=	9
mean		temporal effects		site location effects		interaction effects		total

There are, however, only four factorial treatments (00, 10, 01, 11) and four observations per replicate in this experiment. In order for the estimation problem to be properly posed (nonsingular) for any number of replicates we must reduce the number of independent unknown parameters to four, either by eliminating redundant parameters or by imposing constraints. The most commonly used constraints for this problem are:

$$\sum_{i=1}^{2} \alpha_i = 0 \qquad \sum_{i=1}^{2} \gamma_{ij} = 0 \qquad j = 1,2$$

$$\sum_{j=1}^{2} \beta_j = 0 \qquad \sum_{j=1}^{2} \gamma_{ij} = 0 \qquad i = 1,2 \tag{3-9}$$

These can be used to eliminate five of the nine unknown model parameters so that the resulting minimization problem is nonsingular.

In this special case the least squares estimate minimizing $S(\hat{\theta})$ may be found by setting the derivatives of $S(\hat{\theta})$ with respect to any four independent unknowns equal to zero:

$$\frac{\partial S(\hat{\theta})}{\partial \mu} = 0$$

$$\frac{\partial S(\hat{\theta})}{\partial \alpha_1} = 0$$

$$\frac{\partial S(\hat{\theta})}{\partial \beta_1} = 0$$

$$\frac{\partial S(\hat{\theta})}{\partial \gamma_{11}} = 0$$

The remaining unknowns are removed by applying the constraints (slightly rewritten):

$$\alpha_2 = -\alpha_1$$

$$\beta_2 = -\beta_1$$

$$\gamma_{12} = \gamma_{21} = -\gamma_{11}$$

$$\gamma_{22} = \gamma_{11}$$

The resulting four equations in four unknowns are easily solved using standard numerical methods. Similar principles apply to much more complex problems (see Kendall and Stuart, 1973, Chapter 19).

Constraints can be imposed on minimization problems for reasons other than uniqueness. Suppose, for example, that we wish to estimate the mean μ and temporal effects α_1 and α_2 of the above model *given* the assumption (or hypothesis) that the spatial effects and interactions are all zero:

$$\beta_1 = \beta_2 = 0$$

$$\gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0 \qquad (3\text{-}10)$$

This is equivalent to replacing the original model with a simpler equation which incorporates the constraints imposed by the hypothesis:

$$y_{ijk} = \mu + \alpha_i + e_{ijk}$$

Note that the uniqueness constraints of Equation (3-9) still apply. Those of Equation (3-10) are additional constraints imposed to enable us to investigate the hypothesis that spatial and interaction effects are insignificant.

The quality of a constrained least squares estimate (or of any estimate, for that matter) may be measured by two convenient performance indicators, the estimator bias and covariance defined as:

$$\text{Estimator bias:} \qquad b = E(\hat{\Theta}) - \Theta \tag{3-11}$$

$$\text{Estimator covariance:} \quad \Sigma_{\hat{\Theta}} = E\{[\hat{\Theta} - E(\hat{\Theta})]\,[\hat{\Theta} - E(\hat{\Theta})]^T\} \tag{3-12}$$

Here the expectation operator $E\{\cdot\}$ averages the quantity in brackets over an appropriate probability density. In the case of $\hat{\Theta}$ this implies:

$$E(\hat{\Theta}) = \int_{-\infty}^{\infty} \hat{\Theta} p\,(\hat{\Theta})\,d\hat{\Theta} \tag{3-13}$$

Since Θ and $\hat{\Theta}$ are vectors in the univariate experimental design problem the estimator covariance is a square matrix.

When the residual errors all have the same variance (σ_e^2) and they are uncorrelated with one another, the bias and covariance of the least squares estimator are given by Kendall and Stuart (1973, Chapter 19):

$$b = 0 \tag{3-14}$$

$$\Sigma_{\hat{\Theta}} = \sigma_e^2\, M^{-1} \tag{3-15}$$

where M is a square matrix, commonly called the *Fisher information matrix*, which depends on the observation matrix X and on the constraints imposed as side conditions in the minimization problem. Detailed discussions of the procedures for computing constrained least squares estimates are provided in both Kendall and Stuart (1973, Chapter 19) and Scheffe (1959, Chapters 1-4).

Equations (3-14) and (3-15) indicate that, in the special case of uncorrelated, homoscedastic (equal variance) residual errors, the least squares estimate $\hat{\theta}$ has the following important properties:

1. Values obtained for $\hat{\theta}$ will be distributed evenly about the true effect θ (i.e., the estimate will be unbiased).

2. The range of $\hat{\theta}$ values distributed about θ, as measured by the estimate covariance, will depend only on the residual error variance (σ_e^2) and the Fisher information matrix.

A third important property of the least squares estimator may also be derived (see Kendall and Stuart, 1973, Chapter 19):

3. The estimation variance $\Sigma_{\hat{\theta}}$ associated with $\hat{\theta}$ is the smallest variance attainable with *any* estimator.

These properties, particularly unbiasedness and minimum variance, account for the popularity of least squares estimation in experimental design applications.

Although the above discussion emphasizes estimation of treatment effects, it applies equally well to the population means of interest in sample survey studies. In survey applications, the estimate of most interest is $\hat{\mu}$, the sample mean over all strata, and the other model parameters are included only to properly account for sampling variations from stratum to stratum. The comments made on uniqueness and the need for constraints apply in experimental applications also apply to survey studies. Details on survey estimation are presented in Chapters 39 and 40 of Kendall and Stuart (1976).

3.2.3 Detection of Significant Treatment Effects

Since even the best least squares estimate is likely to differ from the true effect parameter, it is reasonable to ask how this estimate should be interpreted by the ecologist concerned with defining treatment impacts. In controlled experiments the primary question of interest is whether or not different treatments have different effects on species abundance. It is the *relative* magnitude of alternative effects rather than the absolute value of any particular effect which is important. This is a markedly different orientation than that of survey sampling, which focuses on the magnitude of a single parameter, the population mean. Once this mean is estimated, the analysis of survey data is complete.

The emphasis on relative effects and comparisons in controlled experiments usually leads the investigator to postulate a number of hypotheses which may be investigated statistically. Typical hypotheses for the 2x2 example of Equation (3-2) are:

Hypothesis	Corresponding Constraints
H_A: temporal effects are not significant	$\alpha_1 = \alpha_2 = 0$
H_B: site location effects are not significant	$\beta_1 = \beta_2 = 0$
H_{AB}: interactions between time and site locations are not significant	$\gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$

As was noted in the previous section, effect hypotheses may be formulated mathematically as constraints imposed on the least squares minimization problem. Generally, these constraints are simple equations similar to those noted beside each of the hypotheses listed above.

Statistical theory provides a means for testing the validity of hypotheses which can be expressed as estimate constraints. The basic concept is to apply a *decision rule* which states whether or not available information supports the hypothesis of interest. This rule is generally based on a scalar *test statistic* computed directly from field measurements.

Since decision rules are based on test statistics computed from field measurements, they are affected by the measurement uncertainties discussed earlier in conjunction with Figure 3-2. There is always a chance that any decision rule will make an error and either falsely accept or falsely reject a hypothesis. The *decision risk* associated with a decision rule can be quantitatively defined in the following way:

$$R = C_I \alpha + C_{II} \beta \tag{3-16}$$

where

R = decision risk (cost of an incorrect decision)

C_I = cost associated with falsely rejecting the hypothesis (a "type I" error)

α = probability of falsely rejecting the hypothesis (should not be confused with the subscripted effect α_i)

C_{II} = cost associated with falsely accepting the hypothesis (a "type II" error)

β = probability of falsely accepting the hypothesis (should not be confused with the subscripted effect β_j)

Obviously, the "best" decision rule in any given situation is the one which minimizes the decision risk over all possible model parameter values.

In some applications, it is natural to assign a much higher cost to false acceptance of a hypothesis than to false rejection ($C_{II} >> C_I$). This would occur if we were very cautious about saying that *there is no significant difference* in abundance before and after oil is applied to an experimental site. Here the burden of proof is put on the investigator seeking to show no effect. Alternatively, we may wish to be cautious in the opposite direction and assign a much higher cost to false rejection than to false acceptance ($C_I >> C_{II}$). This would occur if we were hesitant about deciding that *there is a significant difference* in abundance before and after oil is applied to the site. Here the burden of proof is put on the investigator seeking to show an effect.

To illustrate the impact of cost assignment on a decision rule, consider two extreme cases:

Case 1: $C_I = 0$ (No penalty assigned to false rejection)
$C_{II} > 0$

Case 2: $C_I > 0$ (No penalty assigned to false acceptance)
$C_{II} = 0$

It is easy to see that the minimum risk in each case (R = 0) is given by decision rules having the following properties:

Case 1: $\beta = 0$ (Always reject the hypothesis)
$\alpha = 1$

Case 2: $\beta = 1$ (Always accept the hypothesis)
$\alpha = 0$

Such extreme rules do not require us to examine the data at all since we always make the same decision! In realistic situations, we will assign non-zero values to both C_I and C_{II} and consequently seek decision rules which provide error probabilities somewhere between 0 and 1.

In order to see how a decision rule is actually implemented and how the type I and type II errors of such a rule are related, it is useful to examine the F-test commonly used to test linear hypotheses such as H_A. This test simply compares the so-called F-statistic for H_A (computation of this statistic is discussed at the end of the section) to a constant:

Accept the hypothesis if: $\hat{F}(Y) < F_\alpha$

Reject the hypothesis if: $\hat{F}(Y) \geq F_\alpha$ (3-17)

The F-statistic $\hat{F}(Y)$ is computed from a particular set of measurements (Y) provided by a particular experimental layout. The decision threshold F_α is selected to provide specified test performance characteristics.

Since $\hat{F}(Y)$ is a random variable by virtue of its dependence on Y, it has a probability density function. The density function differs in shape depending on whether the hypothesis is true or false. The two alternatives for the case of H_A are shown in Figure 3-3a and 3-3b:

Figure 3-3a: H_A true $(\lambda = \alpha_1 - \alpha_2 = 0)$

Figure 3-3b: H_A false $(\lambda = \alpha_1 - \alpha_2 = \lambda_o \neq 0)$

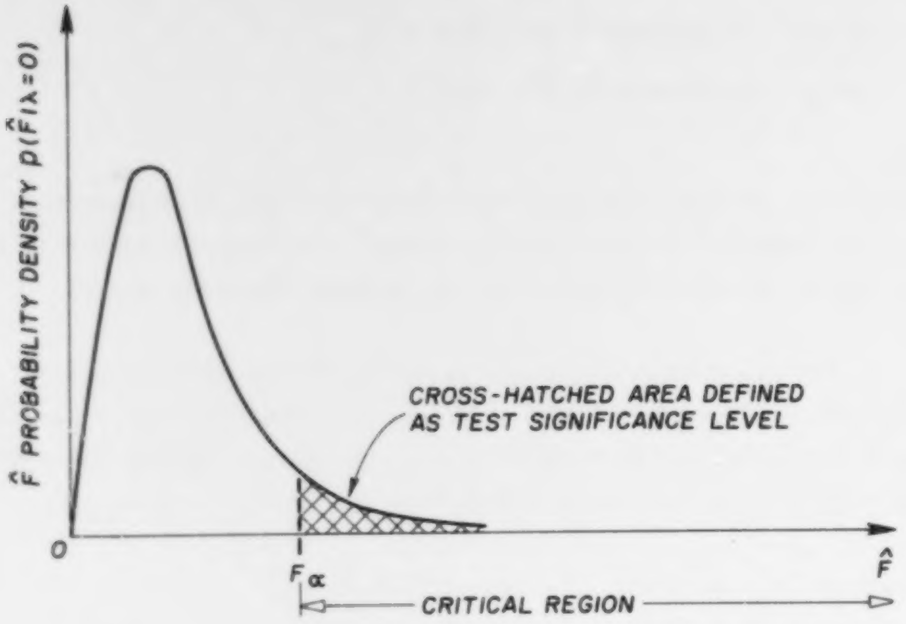
The figure shows that F tends to take on larger values when H_A is false than when H_A is true. The set of values above F_α which result in rejection of the hypothesis (whether it is actually true or not) forms what is known as the *critical region*. In general, the critical region associated with a decision rule is some subset of the "measurement space" defined by the observations. This concept is developed in detail in Chapters 22-24 of Kendall and Stuart (1973).

It is easy to see from the decision rule of Equation (3-16) that the following relationships hold:

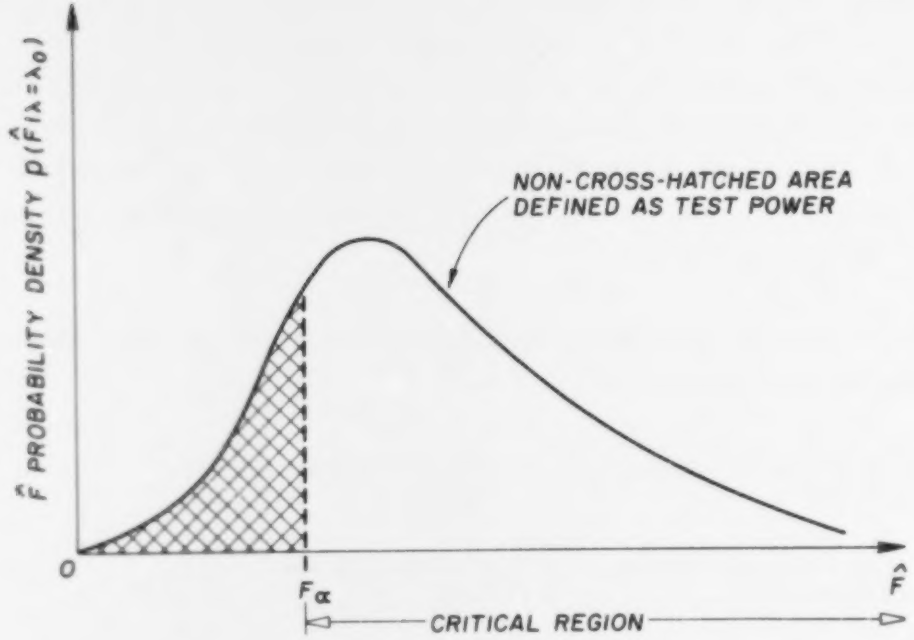
Probability of falsely rejecting H_A = Probability that $\hat{F} \geq F_\alpha$ when $\lambda = \alpha_1 - \alpha_2 = 0$

Probability of falsely accepting H_A = Probability that $\hat{F} < F_\alpha$ when $\lambda = \alpha_1 - \alpha_2 = \lambda_o \neq 0$

The first of these probabilities (α) is given by the cross-hatched area in Figure 3-3a while the second (β) is given by the cross-hatched area in



a) Probability Density of $\hat{F}$ When the General Linear Hypothesis is True ($\lambda=0$)



b) Probability Density of $\hat{F}$ When the General Linear Hypothesis is Not True ($\lambda=\lambda_o\neq 0$)

FIGURE 3-3

PROBABILITY DENSITIES OF THE F STATISTIC USED TO TEST THE GENERAL LINEAR HYPOTHESIS

Figure 3-3b. The common term for the error probability α is the test *significance level*. Since the error probability β is not as commonly used, it does not have a special name. The probability of *correctly* rejecting the hypothesis does have such a name, however. This probability, commonly called the test *power*, is directly related to β through the expression:

$$\text{Test power} = 1 - \beta$$

High test power is equivalent to a low probability (β) of commiting a type II error.

The important point illustrated by Figure 3-3 is that the type I and type II error probabilities (or, equivalently, the significance level and power) of a test are not independent. Changes in the size of the critical region can increase test power only at the expense of also increasing the significance level. Similarly, adjustments which decrease the significance level also decrease test power. The power and significance level of the F-test can both be improved only if the F-statistic's probability density can be made more sharply peaked below F_α when H_A is true and more sharply peaked above F_α when H_A is false. This is, of course, equivalent to a reduction in the test statistic variance. Similar principles apply to any practical hypothesis testing procedure.

With the relationship between power and significance level clarified, we return to our consideration of decision risk. The general expressions for α and β in the case of the F-statistic are given by the cross-hatched regions of Figures 3-3a and 3-3b, respectively. The appropriate mathematical definitions are:

$$\alpha = \int_{F_\alpha}^{\infty} p(\hat{F} \mid \lambda = 0)\, d\hat{F} \qquad (3\text{-}18)$$

$$\beta = \int_{0}^{F_\alpha} p(\hat{F} \mid \lambda = \lambda_o)\, d\hat{F} \qquad (3\text{-}19)$$

If these are substituted into Equation (3-16) the value of F_α which minimizes the decision risk may be computed. As might be expected, this value depends on the costs C_I and C_{II} assigned to the two types of errors.

It is evident from the above that the derivation of a decision rule is relatively straightforward once an appropriate test statistic has been found. This is not always a simple matter--there are no general methods for the derivation of test statistics which give uniformly minimum risk for all model parameter values (see Kendall and Stuart, 1976, Chaps. 22-25 for a detailed discussion). In certain special cases, a constructive method known as the likelihood ratio approach may be used to derive test statistics which have desirable statistical properties. Such a special case arises when the hypotheses of interest lead to linear constraints of the form:

$$A \Theta = C \qquad (3\text{-}20)$$

where

Θ = effects vector (P independent components)

C = constant vector (R components)

A = matrix specifying how the effects are related in each of the constraints (rank P)

The hypotheses H_A, H_B, and H_{AB} defined at the beginning of this section are examples of such *linear hypotheses*. When the constraints of the hypothesis are linear and the residual errors are uncorrelated, zero mean, normally distributed random variables of equal variance σ_e^2, the likelihood ratio approach yields the F-statistic of Figure 3-3. Kendall and Stuart (1976, Chapter 24) show that no other test statistic provides better performance than this one, given the assumptions stated above.

It is important to note that the convenient assumptions used to derive the F-statistic from likelihood ratio theory are not always justified in ecological applications. Even a casual review of available data shows that intertidal abundance measurements frequently have a non-uniform variance structure (variances are typically proportional to mean abundance). Significant correlations and non-normal distributions also occur.

Kendall and Stuart (1973, Chapter 37) discuss a number of techniques which may be used to transform correlated or non-uniform variance data into a form consistent with the assumptions used to derive the F-test. Log transformations such as:

$$y'_{ijk} = \log (y_{ijk})$$

$$\text{or} \qquad y'_{ijk} = \log (y_{ijk} + 1)$$

are frequently used in intertidal applications (see Chapter 4 for an example). These techniques are, however, applicable only when enough historical data are available to indicate whether or not a particular transformation is needed. Such data may be available during the later stages of a sequential research program but they will not generally be available during the initial design phase.

Fortunately, the difficulty of specifying an appropriate transformation in advance is not really important in preliminary design applications. This is because the relative statistical performance of any experimental layout is invariant under a monotonic transformation of variables. A design which maximizes the information content of the y_{ijk} observations collected in a two factor experiment will, for example, also maximize the information content of transformed observations such as log (y_{ijk}). *This allows us to develop experimental designs without regard for the form of the transformations ultimately used in the experimental analysis, so long as the transformation functions are monotonic.*

Studies of the F-test (Kendall and Stuart, 1976, Chap. 37) indicate that its power is relatively insensitive to departures from the underlying assumptions about normality, uniform variance, and correlation which are adopted in the likelihood ratio derivation. As might be expected, this *robustness* property is a desirable trait in any statistical test or estimation procedure. It makes the F-test particularly attractive for impact assessment studies where residual error correlations and variance non-uniformities are common. For all practical purposes, we can confine our review of hypothesis testing procedures to those based on this test.

Extensive discussions of F-statistic computation for commonly encountered hypothesis tests are provided in most books on the analysis of variance--the text by Scheffe (1959) is particularly thorough and informative. Here we will concentrate only on the general principles involved. The fundamental statistical quantities used to compute the F-statistic are the residual error sums-of-squares computed for two different conditions:

Condition assumed	Corresponding residual error sums-of-squares
Hypothesis H_A does not apply	$S_o = (Y - X\hat{\Theta}_o)^T (Y - X\hat{\Theta}_o)$
Hypothesis H_A does apply	$S_A = (Y - X\hat{\Theta}_A)^T (Y - X\hat{\Theta}_A)$

The symbol $\hat{\Theta}_A$ represents the least squares estimate of Θ computed when the linear constraints associated with H_A are imposed as side conditions in the minimization problem (constraints required for uniqueness may also be imposed). The symbol $\hat{\Theta}_o$ represents the least squares estimate computed when these H_A constraints are omitted. Since an unconstrained estimate always gives a better fit to the measured data than a constrained one, the following inequality always holds:

$$S_A \geq S_o$$

Under the assumption that the residuals are normal, uncorrelated random variables of equal variance, the sums of squares S_A and S_o will have chi-squared distributions. The F-statistic simply compares their relative magnitudes:

$$\hat{F} = \frac{(N - P)}{R} \frac{S_A - S_o}{S_o} \qquad (3\text{-}21)$$

where

N = total number of samples (i.e., the dimension of Y)

P = total number of *independent* effects parameters being estimated (i.e., the rank of A)

R = total number of constraints imposed by H_A (i.e., the dimension of the constraint vector c).

The numerator of this expression, commonly called the *treatment sum-of-squares*, may be used to estimate the mean-squared error attributable to treatments. This sample estimate is given by:

$$\hat{\sigma}_t^{\,2} = \frac{S_A - S_o}{R} = \frac{(\hat{\Theta}_A - \hat{\Theta}_o)^T X^T X (\hat{\Theta}_A - \hat{\Theta}_o)}{R} \qquad (3\text{-}22)$$

Similarly, the denominator (or *residual error sum-of-squares*) may be used to estimate the residual mean-squared error, (i.e., the residual error variance). This sample estimate is given by:

$$\hat{\sigma}_e^{\,2} = \frac{S_o}{N - P} = \frac{(Y - X\hat{\Theta}_o)^T (Y - X\hat{\Theta}_o)}{N - P} \qquad (3\text{-}23)$$

These expressions clearly show that the F-statistic is a normalized measure of the euclidean distance between the constrained and unconstrained effects estimates. When the estimates are close, the F-statistic will be small and the hypothesis will be accepted. When the estimates are far apart, the F-statistic will be large and the hypothesis will be rejected.

We have seen in this section that test statistics may be used to help us decide whether or not a particular hypothesis about the effects of factorial treatments should be accepted or rejected. In practice, the hypotheses of most interest will usually be those which compare the effects of different levels of a given factor, to see whether or not this factor contributes significantly to observed sample variability. But other hypotheses will occasionally be of interest also, such as those involving interactions between factors. Whatever the hypothesis being investigated, there will always be a chance that the decision made will be incorrect. The degree of risk associated with an experiment's hypothesis tests will, in general, determine whether or not the experiment will meet an investigator's research needs. If the risk of making an incorrect inference is too high or if the expenditure of effort required to decrease the risk is too great, the experiment cannot be justified statistically. This topic is pursued further in the next section.

3.3 FACTORS INFLUENCING THE PERFORMANCE AND FEASIBILITY OF EXPERIMENTAL DESIGNS

3.3.1 General Design Considerations

The statistical terminology introduced in the previous section allows us to restate in more detail the experimental design objectives briefly mentioned in the introduction to this chapter. *The primary purpose of a controlled experiment is the investigation of the effects of various manipulated or controlled factors on selected dependent variables.* In intertidal applications, the dependent variables of most interest are usually species abundances, diversity indices, or similar quantities which relate in some way to community function and structure. The controlled factors may be physical (tidal elevation, site exposure, etc.), biological (e.g., Katherina abundance) or chemical (petroleum or substrate composition). In order to evaluate the effects of these factors on the species of interest, the investigator measures abundance in various sampling quadrats situated throughout the experimental site in such a way as to permit all factors to be properly controlled and the detrimental effects of nuisance variables to be minimized. *Experimental design is, for the most part, simply the process of defining experimental factors, selecting factor levels (i.e., treatments), and locating sampling quadrats.*

Once the experimental data have been collected, the investigator is faced with the problem of estimating the effects of his treatments and of deciding whether or not these effects differ significantly. These two problems are discussed in detail in Section 3.2. There we point out that statistically valid estimates of treatment effects (or of survey sample population means) must be based on a mathematical model which explicitly relates field observations to experimental factors as well as to potential sources of error. Once this is done, least squares techniques may be used to estimate effects defined in the model and hypothesis testing procedures may be used to determine whether or not these effects are significant from an impact assessment point of view.

Although experimental analysis and design are often approached as distinct problems, they are actually closely related. Obviously, the success of an experimental analysis depends on the quality of data produced by the design. What is somewhat less obvious is the fact that design must anticipate analysis in certain important respects. In particular, the model and hypothesis tests adopted in the experimental analysis define the conceptual framework for the selection of factors, allocation of sampling quadrats and other design decisions made by the experimenter.

Section 3.1 suggests that an optimal experimental design is the one which provides the best information available for a given cost. We can now define "best information" as that set of measurements which minimizes the risk associated with hypotheses investigated during the analysis phase. As we saw in Section 3.2, this risk may be precisely evaluated once the models, hypotheses, test procedures, and decision costs used in the experimental analysis are specified.

It is reasonable to ask here how we can ever find the particular choice of factor levels, sampling quadrats, and experimental procedures which, in fact, minimizes decision risk. Although much effort has been spent on this question (see, for example, Federov, 1972), the only really practical answer is trial and error. We simply must evaluate the risk associated with each plausible design alternative and select the alternative giving the smallest value. An experimenter's candidates for "plausible design alternatives" will depend on his/her understanding of the biological community under investigation, on the resources available to him/her, and on common sense and intuition (if he/she understands the community well enough there may only be one reasonable design alternative). The development of a set of design alternatives is, therefore, a largely subjective and judgmental process which will vary greatly from investigator to investigator. The choice of the best design from this set of alternatives is, on the other hand, a more objective and quantitative process. In this section, we will cover both aspects of design.

A review of the experimental design literature (which is surprisingly qualitative considering the mathematical nature of most statistical literature) reveals a number of general guidelines which should be used to develop candidate experimental designs:

1. Factor selection - It is important that an experiment be constructed to minimize, as much as possible, the influence of unexplained sources of variability or, more precisely, to minimize the residual error variance obtained after the experimental model is fit to field data. This is best achieved by selecting a model (i.e., experimental factors and concomitant variables) which provides a complete description of the environmental phenomena which dominate community behavior. In practice, it is, of course, difficult to know whether or not a model is even marginally adequate until it is tested in the field. This is why a sequential approach to experimentation is essential in most field research studies.

2. Randomization and Blocking - Candidate experimental designs must isolate, as much as possible, effects due to controlled treatments from those due to uncontrolled (or unrecognized) nuisance variables. The sampling layout should be constructed so that extraneous environmental disturbances tend to cancel one another rather than be confused with factorial treatments.

This is most effectively achieved through the process of *randomization*, which we will shortly investigate in detail. In practice, randomization and variance minimization are both closely related to the methods used to group (or block) sampling quadrats at the experimental site. Properly blocked and randomized sampling layouts will accentuate treatment effects and minimize nuisance or extraneous effects.

3. Quadrat Placement - An experimental design must account for logistical or physical constraints which limit the factorial treatment combinations that can be feasibly applied to sampling quadrats at the experimental site. This is particularly important in the case of petroleum studies since oil is difficult to control or even characterize with any accuracy. The sampling quadrats in such studies must be spaced and/or blocked to allow for the uncontrollable spreading of applied oil (the same considerations apply, of course, to studies of accidental spills). Even when experiments do not involve oil, physical constraints may be important. The tidal elevation factor cannot, for example, be randomly allocated to different quadrats since it is by its very nature related to quadrat location. Once a quadrat's coordinates are specified, tidal elevation, substrate type, and other similar factors are implicitly defined. Obviously, such considerations markedly affect the feasibility of candidate designs.

Once a set of plausible and logistically feasible experimental designs has been identified, the performance characteristics of each design can be evaluated statistically. The major design variables of interest are the level of effect considered to be significant by the experimenter, the risk he is willing to take in making inferences about treatment effects, and the number of replicate samples to be collected. If these parameters are specified, the feasibility of any design can be assessed and the best design alternative can be clearly identified. We can now turn to a more detailed consideration of just how this may be done in practice.

3.3.2 Qualitative Aspects of Experimental Design--Factor Selection, Randomization, Blocking, and Quadrat Placement

Factor Selection

As was pointed out in the previous section, the influence of unexplained sources of variability in a controlled experiment is determined, to a large extent, by the validity of the model (i.e., the experimental factors) selected for investigation. An experiment designed to investigate subtle predator-prey interactions at a site dominated by spatial variations in dessication and wave exposure will be doomed to failure unless its design accounts for these environmental factors. It is, of course, easy to come up with a surprisingly long list of physical, chemical and biological variables which *could* affect the local abundance of a species studies in a controlled experiment. The challenge is to select a subset of these factors which explains *most* of the observed abundance variability and yet is still small enough to be manageable.

Systematic methods such as principle component analysis and stepwise regression (Draper and Smith, 1966) are available for identifying the most important environmental factors at a given experimental site. Since these methods require reasonably large quantities of data, they are more relevant to the later stages of a sequential research program than to the initial design phase. Considering the present state of petroleum impact research, and even of intertidal ecosystem research in general, it appears that the dominant environmental factors in impact experiments must be selected judgmentally. Of course, good biological judgment should always be exercised in factor selection, even when principle component analysis or stepwise regression are used.

In order to gain a feeling for how many factors and/or concomitant variables are "manageable" in a given experiment it is helpful to consider how the number of treatments required for each replicate grows as more factors are added. This is illustrated in Figure 3-4, where number of treatments is plotted vs. number of factors for two, three and four level classifications. It is easy to see that hundreds of treatments may be required for even a

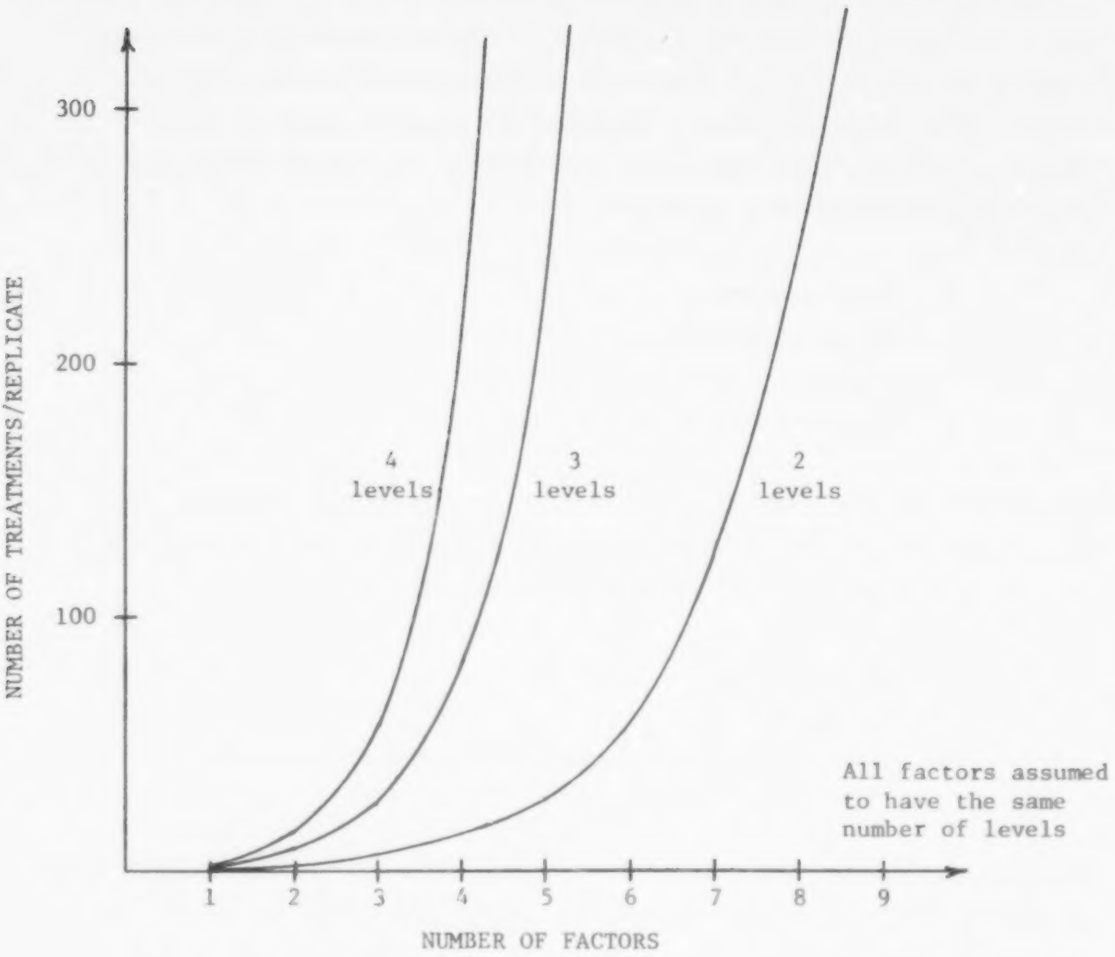


Figure 3-4

Number of Treatments Required in a
Single Factorial Replicate vs.
Number of Experimental Factors

moderate number of factors. Such large treatment numbers pose significant logistic problems as well as statistical difficulties. In intertidal experiments, it may be practically impossible to lay out hundreds of sampling quadrats at a site over an area which is sufficiently homogeneous with respect to nuisance variables. Suppose, for example, that we selected the following factors, each defined at four levels, in a study of Balanus and algal species abundance (see Chapter 4):

1. Tidal elevation
2. Katherina abundance
3. Limpet abundance
4. Carnivorous gastropod abundance

The total number of treatments (quadrats) in this 4^4 experiment is 256. These quadrats would have to be laid out over an area which is relatively uniform with respect to the following nuisance variables:

1. Substrate type
2. Topography
3. Wave exposure (other than that due to differences in tidal elevation)

Furthermore, samples would have to be taken over a sufficiently short period of time to avoid biases or differences in sampling technique casued by changes in personnel, equipment, or weather conditions. An experiment of this size is clearly near the limits of logistic feasibility. If oil were, in addition, added as a factor (either as a result of an intentional perturvation or an accidental spill) an experiment of this size would be unmanageable.

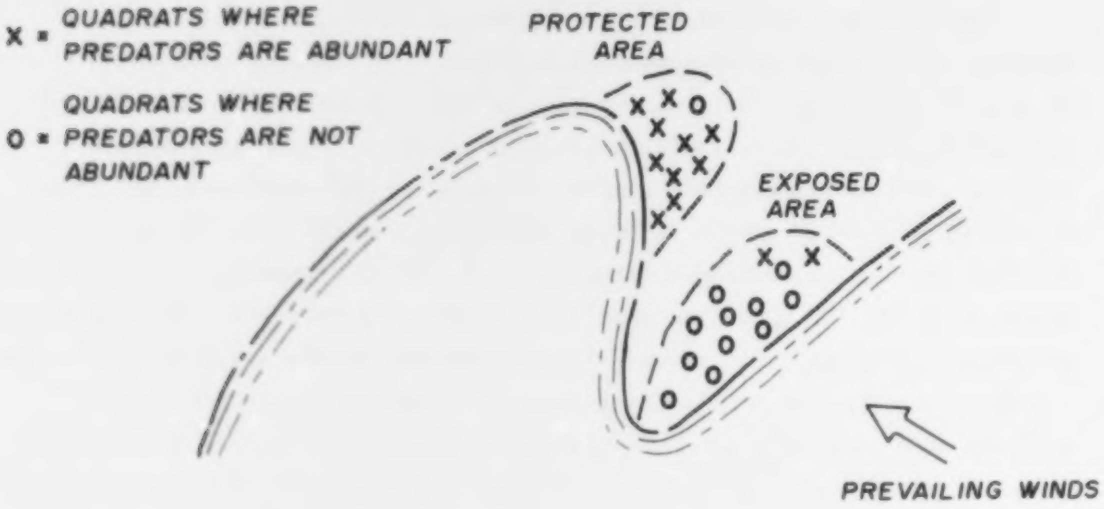
In the case of the above example, the number of required treatments can be substantially reduced (from 256 to 32) without significant loss of information if the four level classifications adopted for the last three factors are replaced by simpler two level (excluded or not excluded) classifications. But even when only two classification levels are used, experiments requiring more than six or eight factors are cumbersome and of dubious usefulness.

Concomitant observations might appear on first glance to offer a solution to the problem of unmanageable numbers of sampling quadrats. If some of the factors originally proposed could be treated as concomitant instead of classified variables, the total number of sampling quadrats could be reduced and the factor still included in the experiment (although in a different form). While this is technically correct, it is important to remember that concomitant observations should not generally be controlled or manipulated. If we did attempt to control concomitants the resulting effects could easily be confused with the effects of the classified factorial treatments applied to the sampling quadrats. Natural fluctuations in a variable such as Katherina abundance may or may not be sufficient to demonstrate a significant effect if this variable is treated as a concomitant. Controlled manipulations (i.e., factorial treatments) of Katherina abundance are more likely to have identifiable effects on the dependent variables of interest. We must, therefore, select concomitant observations with care and realize that they are not necessarily equivalent to controlled experimental factors.

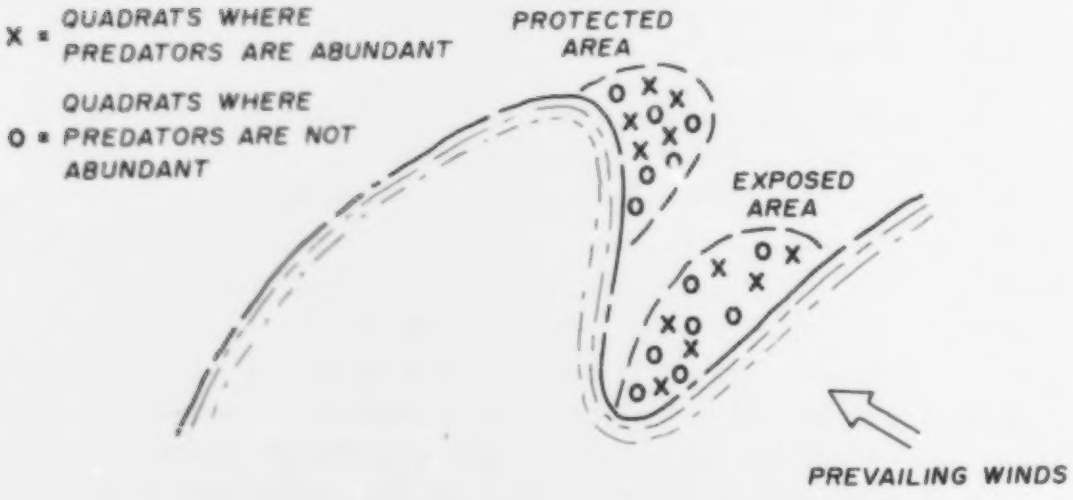
Randomization and Blocking

Proper selection of experimental factors helps to insure that most of the variability observed in field measurements can be explained. But this is not, in itself, enough to guarantee that the experiment's results will be useful or properly interpreted. It is essential that the experimental design also account for and attempt to minimize possible confusion (or "confounding") of effects due to nuisance variables with effects due to experimental factors. As was briefly mentioned in Section 3.2.1, an experiment which includes an interesting biological factor such as predator abundance in its model, but ignores significant environmental variables such as site exposure may falsely attribute exposure-related variations in prey abundance to predator pressure. A hypothetical sampling layout which commits this mistake is shown in Figure 3-5a.

One way to deal with the nuisance variable (site exposure) in this example is to include it as an experimental factor or possibly a concomitant variable in the model. This approach has certain disadvantages, since the


X = QUADRATS WHERE PREDATORS ARE ABUNDANT
O = QUADRATS WHERE PREDATORS ARE NOT ABUNDANT
PROTECTED AREA
EXPOSED AREA
PREVAILING WINDS

a) Non-randomized Sampling Layout for the
Study of Predator Pressure
(predator abundance confounded with exposure)


X = QUADRATS WHERE PREDATORS ARE ABUNDANT
O = QUADRATS WHERE PREDATORS ARE NOT ABUNDANT
PROTECTED AREA
EXPOSED AREA
PREVAILING WINDS

b) Randomized Sampling Layout for the
Study of Predator Pressure
(predator abundance and exposure isolated
by proper randomization)

FIGURE 3-5

Example of the Application of Randomization
to an Experimental Design

number of quadrats or observations required will be increased even though the effects of exposure are not, in themselves, of interest. Besides, nuisance variables are not always as obvious as site exposure is in this example--we cannot reasonably expect to identify *all* potentially important nuisance variables before the experiment is conducted.

An alternative approach is to allocate factorial treatments to sampling quadrats and to locate sampling quadrats on the site in as random a fashion as is possible. This randomization procedure helps to insure that unrecognized nuisance effects will cancel, rather than be confounded with, factorial treatments. A randomized alternative to the layout of Figure 3-5a is shown in Figure 3-5b. Here quadrats are laid out or manipulated (e.g., with cages) so that the predator species is present in both exposed and unexposed quadrats and absent in both exposed and unexposed quadrats. Exposure can no longer be confounded with predator pressure.

In the example of Figure 3-5, quadrats at the exposed and protected sites form two distinct blocks of experimental units. The unrandomized blocks of Figure 3-5a differ in both exposure and predator density while the randomized blocks of Figure 3-5b differ only in exposure (and possibly in other ways not obvious from the figure). Here we see a general principle of experimental design simply illustrated:

> Experimental units (i.e., sampling quadrats) should be grouped into randomized blocks in such a way as to *minimize* nuisance variable variations within each block and to *maximize* nuisance variable variations between blocks.

This practice will not only reduce the chance that nuisance variables will be confounded with experimental factors, it will also generally reduce the residual error variance associated with a given mathematical model. This is because blocking tends to confine within-block variations to treatment-related effects. Between-block variations (called *block effects)* are then due primarily to extraneous error sources, such as nuisance variables, which are of no particular interest to the experimenter.

It is appropriate to mention here that randomized block layouts are not the only statistically legitimate experimental designs found in biological applications. Generally, layout (or treatment allocation) procedures fall into two broad categories:

1. Randomized Block Techniques
    a. Complete block layouts
    b. Incomplete block layouts
        1) Confounded and split-plot designs
        2) Fractionally replicated designs
        3) Combinations of confounding and fractional replication
2. Other Techniques for Grouping Experimental Units
    a. Latin squares
    b. Cross-over designs
    c. Lattice designs
    d. Nested designs

The term *complete block* layout reflects the fact that each possible factorial treatment appears at least once in each of the experimental blocks. *Incomplete block* layouts are simply those which are not complete. To use other terminology, we may say that a complete block contains at least one complete treatment *replicate*, where a replicate is defined as the set of samples associated with all possible factorial treatment combinations--one sample for each treatment. This is the source of the term "fractional replicate" frequently found in discussions of incomplete block designs.

Incomplete block layouts are attractive because they allow an experimenter to investigate some (but not all) treatment effects with a smaller number of quadrats per block than is possible in a complete block layout. This resolves some of the logistic difficulties associated with carrying out multi-factor experiments at sites where nuisance variable nonuniformities are significant. Smaller blocks are generally more uniform and, therefore, more likely to give low residual error variances. The penalty paid for this convenience is the confusion of treatment effects with one another (a phenomenon known as confounding or aliasing, depending on the application). If the confused effects are of little interest or if the confusion can be resolved by reasonable assumptions or external information, this penalty may be insignificant compared to the improvement in experimental performance obtained from the use of smaller blocks.

Unfortunately, the advantages of small block sizes are practically impossible to evaluate before the fact and sometimes even difficult to judge after experimental data have been collected. Consequently, the experimenter must use his best judgment in deciding whether or not within-block nuisance variations are likely to be significant enough to justify an incomplete design. Some general guidelines and comments on the applicability of complete and incomplete randomized block layouts are provided in Appendix B. This Appendix also includes a brief review of the relevance of the other layout procedures listed in category 2 above (latin squares, cross-over designs, etc.). We suggest that the reader refer to Appendix B only after reviewing the rest of Section 3.3.

Quadrat Placement

The practical limitations on quadrat placement and blocking in an intertidal environment are strongly dependent on the particular factors selected for investigation and on the physical and biological characteristics of the experimental site. Although we cannot provide generalized guidelines for quadrat placement, we can identify some of the constraints which will probably be encountered in field studies of petroleum impacts. The specific methods and experimental protocols required to deal with these constraints must be left to the experimenter.

The complete factorial approach requires that every possible combination of factor levels be included in an experiment--each combination defines a factorial treatment to be applied to a distinct sampling quadrat. Randomization theory requires, in addition, that these factorial treatments be randomly allocated to the quadrats in each block. Two practical questions arise as a result of these requirements:

- Can we manipulate the experimental factors at the site in such a way as to obtain all possible factorial combinations or are some combinations impossible due to uncontrollable physical or biological constraints?

- Can the factorial treatments actually be randomly allocated to quadrats or do constraints at the site necessitate more systematic allocation procedures?

Before we design a detailed sampling layout, both of these questions must be answered.

The answer to the first question depends, of course, on the factors selected for study. For most intertidal applications, all necessary factor combinations can be obtained if the experiment is properly defined and sufficient imagination is used in excluding or confining species from particular quadrats. An important exception arises in the case of pervasive treatments such as petroleum. Suppose that we have established that the effect of tidal elevation on the abundance of a species is significant and we wish to examine the impact of petroleum on this elevation-abundance relationship. In principle we can define a 4x2 model of the form:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

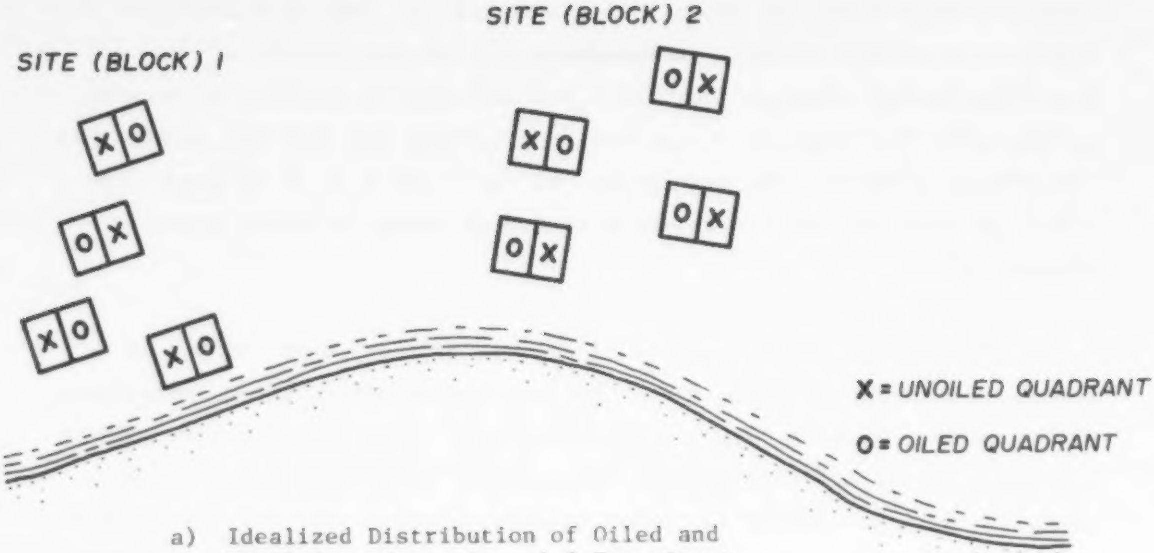
where

α_i = a four level elevation effect

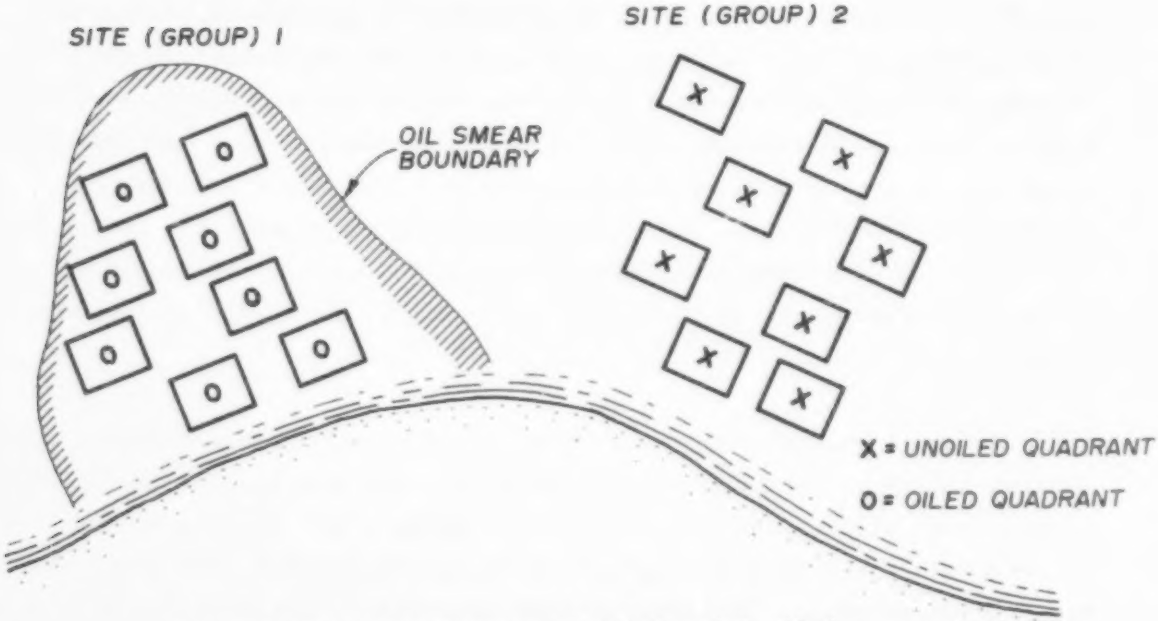
β_j = a two level (present or absent) petroleum effect

γ_{ij} = an elevation-petroleum interaction

This model yields eight factorial treatments, four with petroleum and four without. A theoretically ideal layout for two replicates of this experiment is shown in Figure 3-6a. Unfortunately, once we apply petroleum at the experimental site, it will be dispersed by wave and tidal action and will shortly be spread over all the tidal elevations of interest as well as


SITE (BLOCK) 1
SITE (BLOCK) 2
X = UNOILED QUADRANT
O = OILED QUADRANT

a) Idealized Distribution of Oiled and Unoiled Quadrats in a 4x2 Experiment (two replicates)


SITE (GROUP) 1
SITE (GROUP) 2
OIL SMEAR BOUNDARY
X = UNOILED QUADRANT
O = OILED QUADRANT

b) Realistic Distribution of Oiled and Unoiled Quadrats in a 4x2 Experiment (two replicates)

FIGURE 3-6

Example of Quadrat Location Problems Posed by Petroleum Treatments longitudinally along the shoreline. Consequently, the ideal layout of Figure 3-6a cannot actually be implemented. The very nature of the petroleum factor requires that the oiled and unoiled quadrats be located sufficiently far apart to insure that encroaching oil does not contaminate the control samples. The only practical way to do this is to group the oiled and unoiled quadrats together at two different sites as shown in Figure 3-6b.

A comparison of the layouts of Figures 3-6a and 3-6b reveals an important statistical difference. Whereas each distinct group (replicate) of 8 quadrats shown in Figure 3-6a is a complete randomized block, the two groups of Figure 3-6b are neither complete nor randomized since neither group contains all of the factorial treatments (each contains four rather than eight) and the allocation of treatments to quadrats within each group is not random but systematic (each group contains only one of the oil factor levels). The incompleteness could be eliminated by treating the entire set of 16 quadrats in Figure 3-6b as a single block, but this would probably increase within-block sources of variability due to nuisance variables such as exposure, topography, etc. The resulting residual error variance could then be too large to permit the investigator to detect any statistically significant effects due either to oil or tidal elevation. Here we have a particularly relevant example of practical limitations on quadrat placement imposed by the physical characteristics of one of the experimental factors.

There are several ways to deal with the statistical and logistics problems raised in the example described above. The most obvious one is to drop tidal elevation as a factor and to introduce oil into the problem in a slightly different way. Suppose we retain the two separated sites of Figure 3-6b but measure abundance at both sites before *and* after oil is applied to site 1. We may then postulate the following 2x2 experimental model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where

α_i = a two level temporal effect

β_j = a two level site location (spatial) effect

γ_{ij} = a temporal-spatial interaction

This model and the associated experiment allow us to test for the impact of oil while accounting for other sources of temporal and spatial variability. The inclusion of a site location factor allows us to group the quadrats at each site into a single block without suffering unacceptably large increases in residual error. Statistically speaking, the complete block consists of four samples, one pre-spill and one post-spill at each site. We now have eight replicates instead of two and we can expect considerably more useful information out of our experiment.

Other design procedures are available for the investigator unwilling to give up tidal elevation as a factor. Examples include split-plot designs based on an incomplete block (confounded) analysis and nested designs based on a hierarchal experimental model (see Scheffe, 1959, for more details). In our opinion, these alternatives are rather artificial ways to circumvent a basic physical limitation in the experiment. It is usually better to redefine the experimental objectives and factors than to have to accept the restrictions of incomplete or nonrandomized block designs.

It is worth noting here that the practical design concepts illustrated in this example are relevant to accidental as well as intentional spills. Although pre-spill observations may not always be available, particularly at remote sites, the basic idea of defining factors and blocks to minimize the detrimental effects of nuisance variables still applies. Some experimental and survey sampling models particularly appropriate to accidental spill studies are discussed in Smith (1978). Smith's models and those discussed in this report are, of course, only examples provided to illustrate general principles. Since each experimental situation will pose unique problems and constraints, the experimenter must be prepared to innovate and improvise. There are no substitutes for good judgment.

The largely qualitative considerations discussed in this subsection indicate how a statistically valid and practically feasible experimental layout can be developed. One simply starts with the factors of interest, defines the factorial treatments to be used, and allocates these treatments randomly to appropriate blocks of sampling quadrats. It should be obvious by now that common sense and imagination are important components of any successful design. If no single design is obviously the best, several can be proposed--once the factors and treatments of each alternative are precisely defined, we can evaluate the risk associated with each and choose the one which performs the best. The computation of risk and the role of such variables as number of factor levels, number of replicates, and residual error variance are discussed in detail in the next subsection.

3.3.3 Quantitative Aspects of Experimental Design--Decision Thresholds, Design Parameters, and Feasibility Assessment

In order to understand how the risk associated with an experimental design may be evaluated, we must return to some of the basic hypothesis testing concepts presented in Section 3.2.3. Decision risk was defined there (Equation 3-16) as:

$$\mathcal{R} = C_I\alpha + C_{II}\beta$$

where

C_I = cost associated with falsely rejecting a hypothesis

C_{II} = cost associated with falsely accepting a hypothesis

α, β = error probabilities dependent on the particular decision rule selected for testing the hypothesis

The decision rule most commonly used to test the hypotheses of factorial experiments is based on the F-statistic (see Equation 3-17):

Accept the hypothesis if: $\hat{F}(Y) < F_\alpha$

Reject the hypothesis if: $\hat{F}(Y) \geq F_\alpha$

where the F-statistic $\hat{F}(Y)$ is computed from appropriate residual error sums of squares (the details are provided in Equations (3-21) through (3-23)). The decision threshold F_α defines the boundary of the test's critical region and is selected to provide specified statistical characteristics.

If the F-test is used to test experimental hypotheses, the type I and II error probabilities (α and β) are given by Equations (3-18) and (3-19):

$$\alpha = \int_{F_\alpha}^{\infty} p(\hat{F}|\lambda = 0)d\hat{F}$$

$$\beta = \int_{o}^{F_\alpha} p(\hat{F}|\lambda = \lambda_o)d\hat{F}$$

The *non-centrality parameter* λ appearing in these expressions is a measure of the normalized euclidean distance between actual treatment effects and those which would apply *if* the hypothesis under investigation were true. The general expression for λ is given by (see Kendall and Stuart, 1973, Chap. 24):

$$\lambda = \frac{\sigma_t^2}{\sigma_e^2} \tag{3-24}$$

where

$\sigma_t^2 = \frac{(\Theta_A - \Theta_o)^T X^T X(\Theta_A - \Theta_o)}{R}$ = the true mean-squared error due to treatments

Θ_o, Θ_A = the actual and hypothesized treatment effects, respectively

X = the observation matrix of the experimental model (see below)

σ_e^2 = the true residual mean-squared error

Note that σ_t^2 is the expected value of the sample mean-squared error due to treatments (see Equation 3-22).

The experimental model referred to here is the general linear model given in Equation (3-5):

$$Y = X\Theta + e$$

where

Y = vector of N observations

Θ = vector of unknown treatment effects (P of these are independent)

X = matrix relating treatment effects to observations (rank P)

e = vector of N residual errors

The hypothesized treatment effects Θ_A used in the definition of σ_t^2 obey a set of R linear constraints given by the matrix equation

$$A\Theta_A = c$$

where

c = a vector of R specified constants

A = a matrix specifying how the treatment effects are related in each of the R constraints imposed by hypotheses H_A (rank P)

The practical significance of such constraints is examined later in this section.

With these preliminaries taken care of, we can clearly identify the design parameters which affect the risk associated with an F-test of any linear hypothesis H_A (see Figure 3-7). The risk expression depends directly on C_I, C_{II}, α and β. The error probabilities α and β depend, in turn, on F_α and the central and noncentral F-probability densities $P(\hat{F}|\lambda = 0)$ and $P(\hat{F}|\lambda = \lambda_o)$. These densities depend, finally, on the parameters N-P, R,

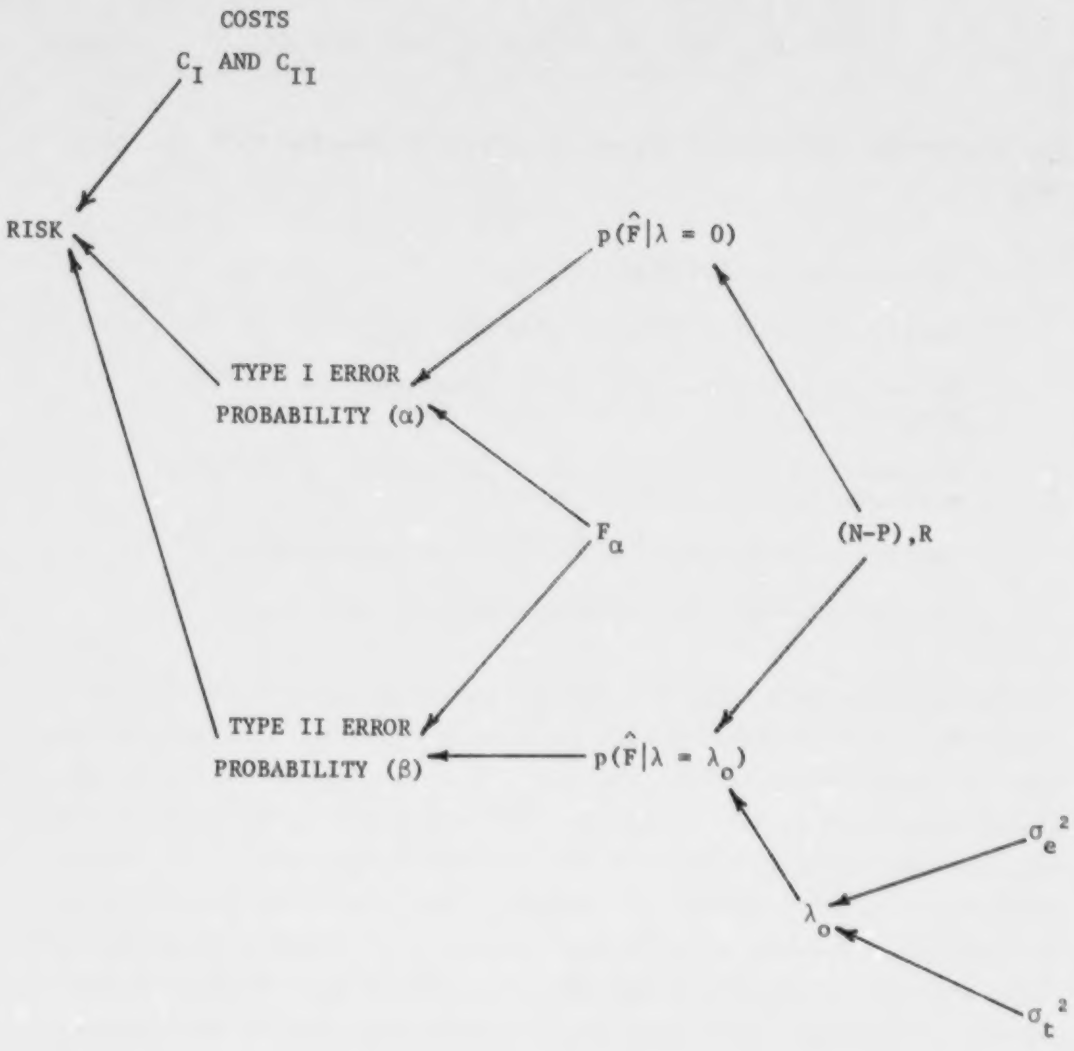

COSTS
C_I AND C_{II}
RISK
$p(\hat{F}|\lambda = 0)$
TYPE I ERROR
PROBABILITY (α)
F_α
(N-P),R
TYPE II ERROR
PROBABILITY (β)
$p(\hat{F}|\lambda = \lambda_o)$
σ_e^2
λ_o
σ_t^2

Figure 3-7

Design Parameters Affecting the Risk
Associated with an F-test of the
Linear Hypothesis H_A

σ_t^2, and σ_e^2. The general functional expression for the normalized decision risk is then:

$$\frac{R}{C_I} = \alpha(F_\alpha, N\text{-}P, R) + \frac{C_{II}}{C_I}\beta(F_\alpha, N\text{-}P, R, \sigma_t/\sigma_e) \qquad (3\text{-}25)$$

The fundamental independent variables governing decision risk are therefore:

- The relative penalty cost (C_{II}/C_I)
- The critical region decision threshold (F_α)
- The number of degrees of freedom (independent variables) associated with the residual error (N-P)
- The number of degrees of freedom (independent constraints) associated with the hypothesis (R)
- The true root-mean-squared error due to treatments (σ_t)
- The true residual root-mean-squared error (σ_e)

In the following paragraphs we briefly discuss alternative methods for selecting F_α and for dealing with the relative cost C_{II}/C_I. We then examine the relationship between the sampling layout itself and the primary design parameters R, N-P, σ_t and σ_e. Finally we review the factors affecting experimental feasibility and consider some useful ways of analyzing experimental sensitivity and performance. For simplicity, we concentrate on randomized complete block designs and leave the details of incomplete block analyses to Appendix B and the cited references. Complete block layouts are probably sufficient for most intertidal impact experiments, particularly for those using less than four or five factors. The complete block analyses presented here illustrate most of the important statistical concepts needed in practical experimental design.

Determination of the Decision Threshold

Since the type I and type II error probabilities appearing in the definition of decision risk are both strongly dependent on F_α it is reasonable to ask how this constant should be determined and how it affects risk. To simplify the analysis, assume that the design and hypothesis have both been specified so that the parameters N-P, σ_e^2, σ_t^2, and R are fixed. In this case, the normalized risk is a function of only C_{II}/C_I and F_α:

$$\frac{R}{C_I} = \alpha(F_\alpha) + \frac{C_{II}}{C_I} \beta(F_\alpha) \tag{3-26}$$

The one-to-one relationship between $\alpha(F_\alpha)$ and $\beta(F_\alpha)$ is discussed in Section 3.2.3 and graphically illustrated in Figure 3-8. A plot of $\beta(F_\alpha)$ vs. $\alpha(F_\alpha)$ for a particular hypothesis and $4x2^3$ sampling layout developed in Section 4.4.2 is shown in Figure 3-8. Such plots may be used to evaluate normalized risk as a function of α for various cost alternatives. Figure 3-9 shows three alternative risk curves based on the specific case shown in Figure 3-8. The unique minimum associated with each cost alternative defines the value of α, and consequently the value of F_α, giving the lowest risk. Note that for this *particular* case, the conventional choice of $\alpha = .05$ corresponds to $C_{II}/C_I = .5$, i.e., the relative cost of rejecting a significant effect (type II error) is one-half the cost of falsely detecting a significent effect (type I error).

This fairly straightforward way of obtaining F_α requires the experimenter to select costs which reflect his best assessment of the relative penalties imposed by type I and type II errors. Figure 3-9 shows that these costs can have a strong influence on α and F_α and can, therefore, determine the practical outcome of the experiment. Unfortunately, there are few, if any, objective measures which can be used to judge whether or not a particular set of costs is reasonable.

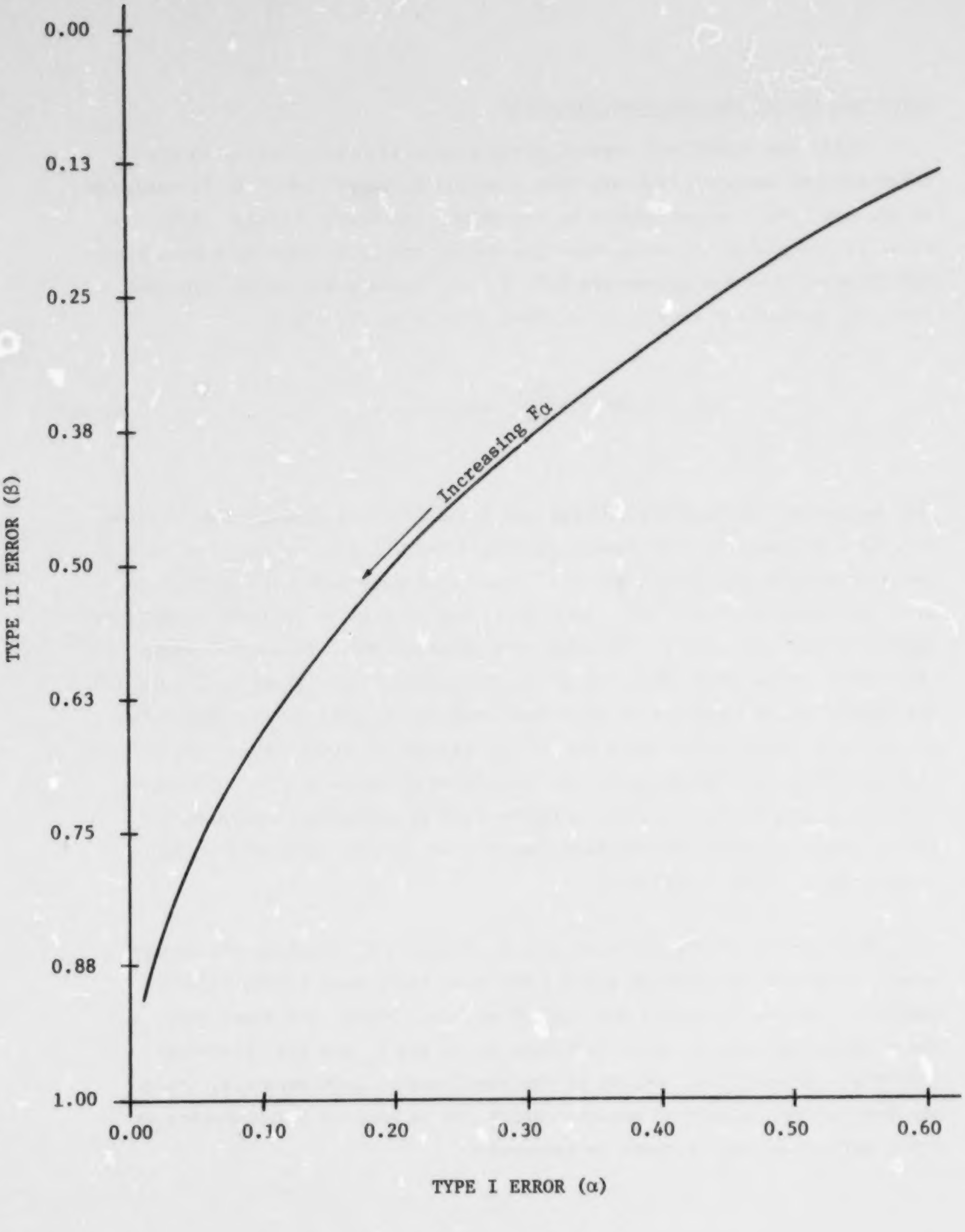


Figure 3-8

Plot of Type II vs. Type I Error Probabilities

for a Particular Experimental Layout

(adapted from Figure 4-6b)

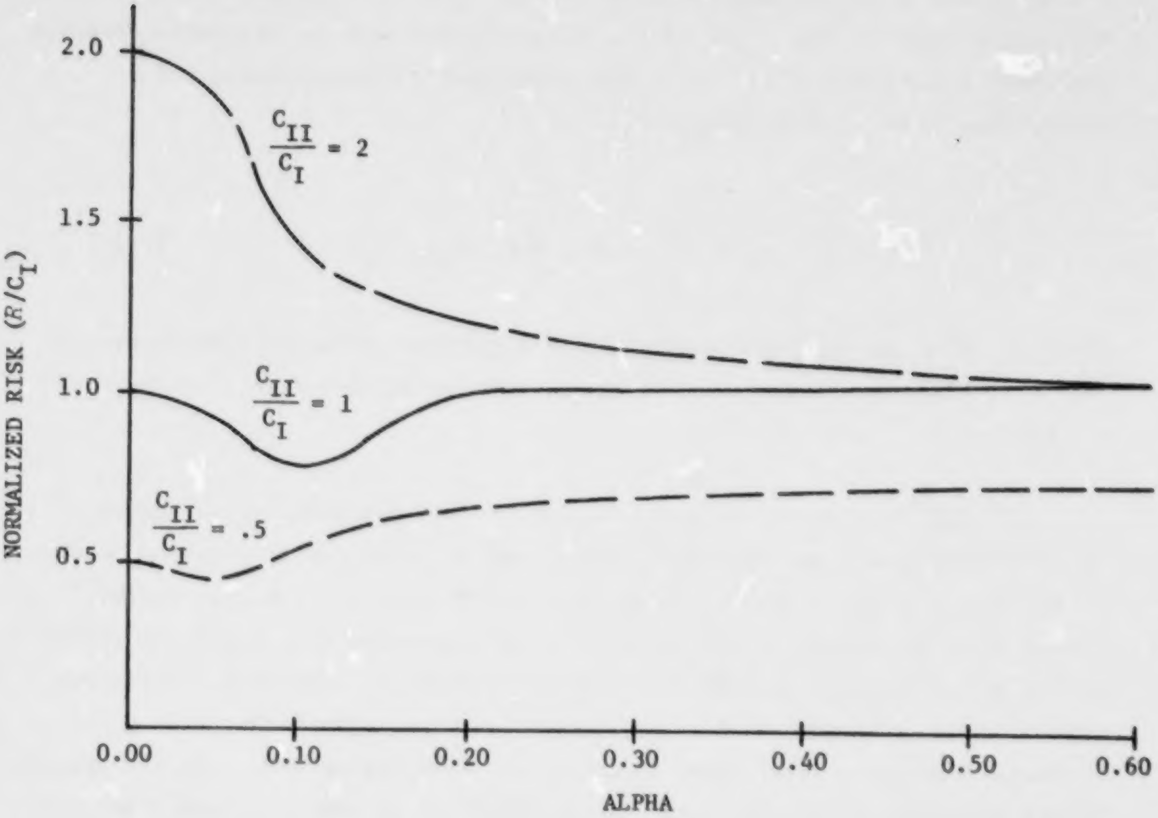


Figure 3-9

Plot of Normalized Risk vs. Significance Level
for Various Relative Cost Alternatives

The difficulties associated with penalty cost specification have led statisticians to seek more objective methods for defining decision thresholds. The most common procedure, developed by Neyman and Pearson and described in Kendall and Stuart (1973, Chapter 24), is to set α equal to a "reasonable" value such as .01, .05, or .10 (the most common alternatives). The specified α (α_o) determines a particular F_α without requiring any consideration of cost or risk. One of the primary advantages of the Neyman-Pearson approach is its simplification of the normalized risk expression for a generalized layout and hypothesis:

$$\frac{R}{C_I} = \alpha_o + \frac{C_{II}}{C_I} \beta(\alpha_o, N\text{-}P, R, \sigma_t/\sigma_e) \qquad (3\text{-}27)$$

Since α_o is a constant, the design which minimizes the type II error probability is also the design which minimizes the normalized risk, *irrespective of the costs* C_I *and* C_{II}.

Although this report uses the Neyman-Pearson approach for selecting decision thresholds, we feel it is important to recognize that this procedure is ultimately just as subjective as cost specification. The experimenter should consider carefully the value of α_o most appropriate to a given application and not simply use .05 or .10 because everyone else does. Often the type II error performance of a test can be improved significantly if α_o is increased above the usual upper limit of .10 (see Figure 3-8). In the absence of any bias one way or the other, one approach is to make the type I and type II error probabilities approximately equal. Certainly, a balanced emphasis on both types of error seems the approach most appropriate in impact assessment studies where either faulty rejection or faulty acceptance of the impact hypothesis can have important policy implications.

Design Parameters Related to the Sampling Layout

Equation (3-27) demonstrates that the Neyman-Pearson approach to threshold specification makes risk minimization functionally equivalent to minimization of the type II error probability function $\beta(\alpha_o, N\text{-}P, R, \sigma_t/\sigma_e)$.

This function is, in fact, the primary design-related measure of experimental performance. Once it is determined, the normalized risk may be evaluated straightforwardly from α_o and C_{II}/C_I, parameters not related to the sampling layout but only to the decision rule adopted.

Although the type II error probability is the fundamental measure of design performance, it is generally of less interest than the test power, which is related to it through the transformation:

$$\text{Power} = 1 - \beta(\alpha_o, N\text{-}P, R, \sigma_t/\sigma_e) \tag{3-28}$$

The test power, as we saw in Section 3.2.3, is simply the probability that the hypothesis under consideration will be correctly rejected if it is false. In impact assessment applications, this is equivalent to the probability that a given experimental factor is judged to have a significant effect on abundance.

If we focus our attention on power, it is evident that the important design parameters are N-P, R, σ_t, and σ_e. In this subsection, we discuss the relationship between these parameters, the sampling hypotheses under investigation, and the structure of the sampling layout. It is easiest to begin by presenting a somewhat more detailed version of the experimental model introduced in Equation (3-5) of Section 3.2.1. Suppose that we partition this model into several components as follows:

$$\underset{\substack{\text{observation}\\ \text{vector}}}{Y} = \underset{\text{mean}}{\mu} + \underset{\substack{\text{main}\\ \text{effects}}}{X_M\Theta^M} + \underset{\substack{\text{interaction}\\ \text{effects}}}{X_I\Theta^I} + \underset{\substack{\text{block}\\ \text{effects}}}{X_B\Theta^B} + \underset{\substack{\text{residual}\\ \text{error}}}{e} \tag{3-29}$$

Each component of this equation contributes a certain number of independent parameters (or degrees of freedom) to the least-squares estimation problem. Kendall and Stuart (1973, Chap. 35) show that, if the appropriate uniqueness constraints are applied, the degree of freedom allocations for the general complete block factorial model take the form indicated in Table 3-2. Modifications to account for incomplete block designs are discussed in

Table 3-2

DEGREE OF FREEDOM ALLOCATIONS
FOR THE GENERAL COMPLETE BLOCK FACTORIAL MODEL

Parameter	Degrees of Freedom
Main effects	
factor 1	I_1-1
factor 2	I_2-1
.	.
.	.
.	.
factor F	I_F-1
Interactions	
factors 1&2	$(I_1-1)(I_2-1)$
factors 1&3	$(I_1-1)(I_3-1)$
.	.
.	.
.	.
factors 1 through F	$(I_1-1)(I_2-1)\ldots(I_F-1)$
Treatment effects subtotal	T-1
Block effects	L-1
Residual error	(K-1)T - (L-1)
Grand mean	1
Other parameters subtotal	(K-1)T + 1
Total number of samples	N = KT

Variable definitions: I_i = number of levels defined for factor i
F = number of factors
L = number of complete blocks
K = number of replicates distributed evently amont L blocks
T = total number of treatments = $\prod_{i=1}^{F} I_i$

Appendix B and in Scheffe (1959, Chap. 5). Since the total number of observations defined by the experimental model is KT and the total number of independent parameters is T + (L-1), the residual error degrees of freedom parameter (N-P) is found by subtraction to be (K-1)T - (L-1). In a complete block experiment (N-P) depends on the number of treatments, replicates, and blocks used in the sampling layout.

Sections 3.2.2 and 3.2.3 note that experimental hypotheses can be described mathematically by an appropriate set of treatment effect constraints. The number of degrees of freedom for each hypothesis (R) is simply the number of independent constraints this hypothesis generates. The hypotheses of most interest in impact assessment applications are those which specify that particular experimental factors or factor interactions have no significant effects on abundance. Such hypotheses lead to constraints which force all appropriate effects parameters to be zero. In the two-factor case, these constraints take the form shown below:

Hypothesis	Corresponding Constraints
H_A: factor A effects are not significant	$\alpha_i = 0 \; ; \; i = 1, I_A$ (I_A constraints)
H_B: factor B effects are not significant	$\beta_i = 0 \; ; \; i = 1, I_B$ (I_B constraints)
H_{AB}: interactions between factors A and B are not significant	$\gamma_{ij} = 0 \; ; \; i = 1, I_A$ $j = 1, I_B$ ($I_A I_B$ constraints)

Uniqueness considerations reduce the number of independent parameters, and therefore the number of independent constraints, associated with each hypothesis. In the case of hypothesis H_A, for example, the uniqueness requirement:

$$\sum_{i=1}^{I_A} \alpha_i = 0$$

insures that if I_A-1 of the α_i's are equal to zero, the remaining α_i will also equal zero. Consequently, hypothesis H_A actually generates only I_A-1 independent constraints. Similar reasoning may be applied to the general complete block model to show that the number of independent constraints associated with each impact hypothesis is equal to the number of independent parameters (degrees of freedom) defined for the corresponding factor or interaction. The degrees of freedom parameter for a no input hypothesis associated with a particular main effect or interaction is therefore simply the value listed in the appropriate row of Table 3-2. In a complete block experiment, R depends only on the number of levels specified for each of the factors included in the hypothesis of interest. Modifications for incomplete designs are discussed in Appendix B and in Scheffe (1959, Chapter 5).

The above analysis defines the design parameters (N-P) and R for a particular sampling layout. We must now consider how the mean-squared errors σ_t^2 and σ_e^2 can be determined. It was pointed out at the beginning of this section that σ_t^2 is a measure of the euclidean distance between constrained (hypothesized) and unconstrained (actual) effects parameters:

$$\sigma_t^2 = \frac{(\Theta_A - \Theta_o)^T X^T X(\Theta_A - \Theta_o)}{R} \qquad (3\text{-}30)$$

When the actual effects (Θ_o) are equal to the hypothsized effects (Θ_A) the hypothesis is true and σ_t^2 is zero. When the actual and hypothesized effects differ, σ_t^2 will be a positive non-zero number.

We do not, of course, know the actual effects applying in a given situation but we can *assume* a set of effects and evaluate the performance (power or risk) that would be obtained *if* these assumed effects actually occurred. In order to keep the performance evaluation conservative, it is wise to assume a worst-case set of actual effects--i.e., to assume that the effects occur in the combination which is most difficult to detect. Scheffe (1959, Chapter 3) shows that this worst case of effects is defined by:

$$\Theta_i - \Theta_j = \Delta \quad ; \quad \Theta_k = (\Theta_i + \Theta_j)/2 \qquad \text{for all } k \neq i \text{ or } j \quad (3\text{-}31)$$

That is, two effects associated with the hypothesis of interest differ by Δ while all other effects are equal to the mean of these two. For a maximum difference in effects equal to Δ, this arrangement gives the lowest test power and highest risk.

If the appropriate version of Equation (3-31) is substituted into Equation (3-30) the worst case values of σ_t^2 for the general F-factor complete block experiment may be derived. These values and the corresponding non-centrality parameters take the general form:

Hypothesis	σ_t^2	Non-centrality parameter (λ)
Main effects	$\frac{N\Delta^2}{2I_i}$	$\frac{N}{2I_i}\left(\frac{\Delta}{\sigma_e}\right)^2$
First-order interactions	$\frac{N\Delta^2}{2I_iI_j}$	$\frac{N}{2I_iI_j}\left(\frac{\Delta}{\sigma_e}\right)^2$
$\vdots$		
(m-1)-order interactions	$\frac{N\Delta^2}{2I_iI_j\ldots I_m}$	$\frac{N}{2I_iI_j\ldots I_m}\left(\frac{\Delta}{\sigma_e}\right)^2$

where $N = K \prod_{i=1}^{F} I_i$ = total number of observations

I_i = number of levels specified for factor i

K = number of replicates

Δ = assumed maximum deviation between treatment effects

σ_e = true residual root-mean-squared error

When the experimental layout is incomplete, these expressions must be modified as discussed in Appendix B.

Since the power and risk functions discussed earlier depend directly on the non-centrality parameter λ and only indirectly on σ_t^2 and σ_e^2, performance evaluations can be based on the normalized variable Δ/σ_e appearing in the expressions for λ given above. This conveniently circumvents the need to estimate σ_e^2 and allows us, for example, to compute the probability of detecting an effect equal to 0.5, 1.0, or 2.0 residual error standard deviations. Although such relative performance evaluations can be used to compare the effectiveness of alternative experimental designs, this approach may be misleading. Suppose, for example, that a particular sampling layout gives a power of .95 for a Δ/σ_e of 0.5. It is tempting to say that an alternative layout giving a power of .70 for the same Δ/σ_e is inferior. This is true only if the two layouts result in comparable values of σ_e. If the second layout has a much lower residual error standard deviation, it will actually detect smaller effects than the first layout.

The basic problem here is that the residual error standard deviation is highly sensitive to the structure of the experiment--to the factors selected, to the location, configuration and size of the sampling blocks, to the distribution of the quadrats, and to the sampling procedures used. It is unrealistic to assume that σ_e will be the same for two different designs except in certain special cases. The most important special case is a comparison between two designs which are identical except for the number of replications taken. Replication does not change the fundamental nature of the layout (unless it requires the addition of more sampling blocks distributed over a larger area at the experimental site) and so does not affect the residual error variance. It does, of course, affect the accuracy of the least-squares effects estimates as well as the test power (through the parameters N-P, R, and σ_t^2).

We are confronted, then, with one of the basic delimmas of practical experimental design--*how can we estimate the residual error variance associated with a particular design before that design is actually tested in the field?* There are really only two answers to this question. We can either assume a reasonable value for residual error variance or we can attempt to estimate this variance from data collected in previous field studies which are at least roughly comparable. In most cases, the latter approach is preferable even though admittedly imperfect.

Most available intertidal field data have been collected in sample survey studies which do not have the same objectives as the controlled factorial experiments of primary interest here. Although we can usually calculate the mean-squared observation error from these data sets, it is difficult to say how much this error could have been decreased if a factorial model had been fit to the data. The situation is clearly demonstrated if we inspect the appropriate model equations. In a simple, unstratified survey (the most frequent type), the model used is of the form:

$$y_i = \mu + e_i$$

A complete two factor experimental model is described by the familiar equation:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

The residual mean-squared error estimates for these two models are given by:

$$\text{Survey } \sigma_e^2: \quad \hat{\sigma}_{es}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2$$

$$\text{Experiment } \sigma_e^2: \quad \hat{\sigma}_{ee}^2 = \frac{1}{N} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{K} (y_{ijk} - \mu - \alpha_i - \beta_i - \gamma_{ij})^2$$

If the effects postulated in the experimental model turn out to be significant σ_{ee}^2 will probably be much smaller than σ_{es}^2 because the experiment explains more of the observed variation than the survey does. We can therefore generally expect residual error variance estimates obtained from sample survey data to be conservatively high.

In some cases, auxiliary data taken during a survey (such as tidal elevation or substrate composition) may be used to fit a simple after-the-fact experimental model to the survey data. If tidal elevation were available, a single factor model would provide the following σ_e^2 estimate:

$$\text{After the fact experiment } \hat{\sigma}_e^2: \quad \sigma_{ea}^2 = \frac{1}{N} \sum_{i=1}^{I_1} \sum_{k=1}^{K} (y_{ik} - \mu - \alpha_i)$$

We can expect this variance estimate to lie somewhere between σ_{es}^2, the survey-based estimate, and σ_{ee}^2, the unavailable estimate that would be obtained from the actual experiment. An interesting example of this approach to residual error variance is discussed in Chapter 4.

If we prefer not to rely on inadequate historical data to develop an estimate for σ_e^2 we can carry out a preliminary experiment with the primary objective of estimating this unknown variance. This approach is practically mandatory in unprecedented or highly uncertain situations such as those involving controlled oil spills. The concept of sequential experimentation mentioned in Section 3.1 is, of course, quite relevant here. Although the performance of the initial survey in a sequential field program cannot be adequately estimated because of uncertainty in σ_e^2, the next experiment in the program will be able to use improved σ_e^2 estimates based on this survey. A sequential research program designed to provide increasingly accurate estimates of experimental performance is discussed in detail in Chapter 4.

Feasibility Assessment

The previous subsection outlines how a particular experimental design can be characterized by a simple set of statistical variables--the degrees of freedom for residual error (N-P) and for the hypothesis of interest (R), the assumed maximum treatment effect deviation (Δ), and the residual error standard deviation (σ_e). The expressions derived there are conveniently summarized in the analysis of variance (ANOVA) table for a general complete block experiment presented in Table 3.3. These expressions may be evaluated manually or on a computer for any specified layout and the parameters $\nu_1 = R$, $\nu_2 = N-P$, α_o and λ substituted into appropriate expressions for the non-central F probability distribution:

$$\text{power} = \int_{F_\alpha(\alpha_o)}^{\infty} p(\hat{F}|\nu_1,\nu_2,\lambda)\,d\hat{F}$$

Useful approximations for this integral are given in Kendall and Stuart (1973, Chap. 24).

Table 3-3

ANALYSIS OF VARIANCE TABLE FOR THE GENERAL COMPLETE BLOCK FACTORIAL EXPERIMENT

Hypothesis	Constraint	Degrees of Freedom	Non-Centrality Parameter
Main effects			
H_i: factor i significant	$\alpha_i=0$ $(i=1,\ I_i-1)$	I_i-1	$\frac{N}{2I_i}\left(\frac{\Delta}{\sigma_e}\right)^2$
First-order interactions			
H_{ij}: interactions between factors i and j are significant	$\gamma_{ij}=0$ $(i=1,\ I_i-1$ $j=1,\ I_j-1)$	$(I_i-1)(I_j-1)$	$\frac{N}{2I_iI_j}\left(\frac{\Delta}{\sigma_e}\right)^2$
(m-1)-order interactions			
$H_{ij\ldots m}$: interactions between factors i,j,...m are significant	$\gamma_{ij\ldots m}=0$ $(i=1,\ I_i-1$ $j=1,\ I_j-1$ $m=1,\ I_m-1)$	$(I_i-1)(I_j-1)\ldots(I_m-1)$	$\frac{N}{2I_iI_j\ldots I_m}\left(\frac{\Delta}{\sigma_e}\right)^2$
Block effects	--	(L-1)	--
Residual error	--	(K-1)T - (L-1)	--
Grand mean	--	1	--
Total number of samples		N = KT	

Variable definitions:

I_i = number of levels defined for factor i

T = total number of treatments = $\prod_{i=1}^{F} I_i$

F = number of factors

L = number of complete blocks

K = number of replicates distributed evenly among L blocks

Δ = assumed maximum deviation between treatment effects

σ_e = true residual root-mean-squared error

α_i, γ_{ij} = treatment effects

Our ability to evaluate power (and therefore risk) as a function of at least seven or eight distinct design parameters presents us with a potentially large number of parametric comparisons--power vs. replicates for various σ_e^2 values, power vs. Δ for various α_o values, etc. Nearly all of these comparisons are useful for one purpose or another as shown by the numerous examples of parametric performance comparisons provided in Chapter 4 (these were generated by RMA's experimental design evaluation program).

It is useful to review the factors which determine the *feasibility* of a practical experiment and to identify a few performance comparisons which are particularly relevant to feasibility assessment. *A feasible impact experiment is one which can detect significant ecological effects (e.g., abundance changes) with acceptable risk performance for an acceptable expenditure of effort.* This definition obviously requires that the experimenter specify rather precisely the following:

- The level of effect considered to be significant (Δ_s).
- The maximum risk, or, equivalently, the power and significance level considered to be acceptable.
- The maximum expenditure of effort--number of replicates or samples, number of treatments, temporal and spatial scope of the experiment--considered to be acceptable.

In addition, the experimenter must estimate the residual error variance which can reasonably by expected to result from the experiment under consideration.

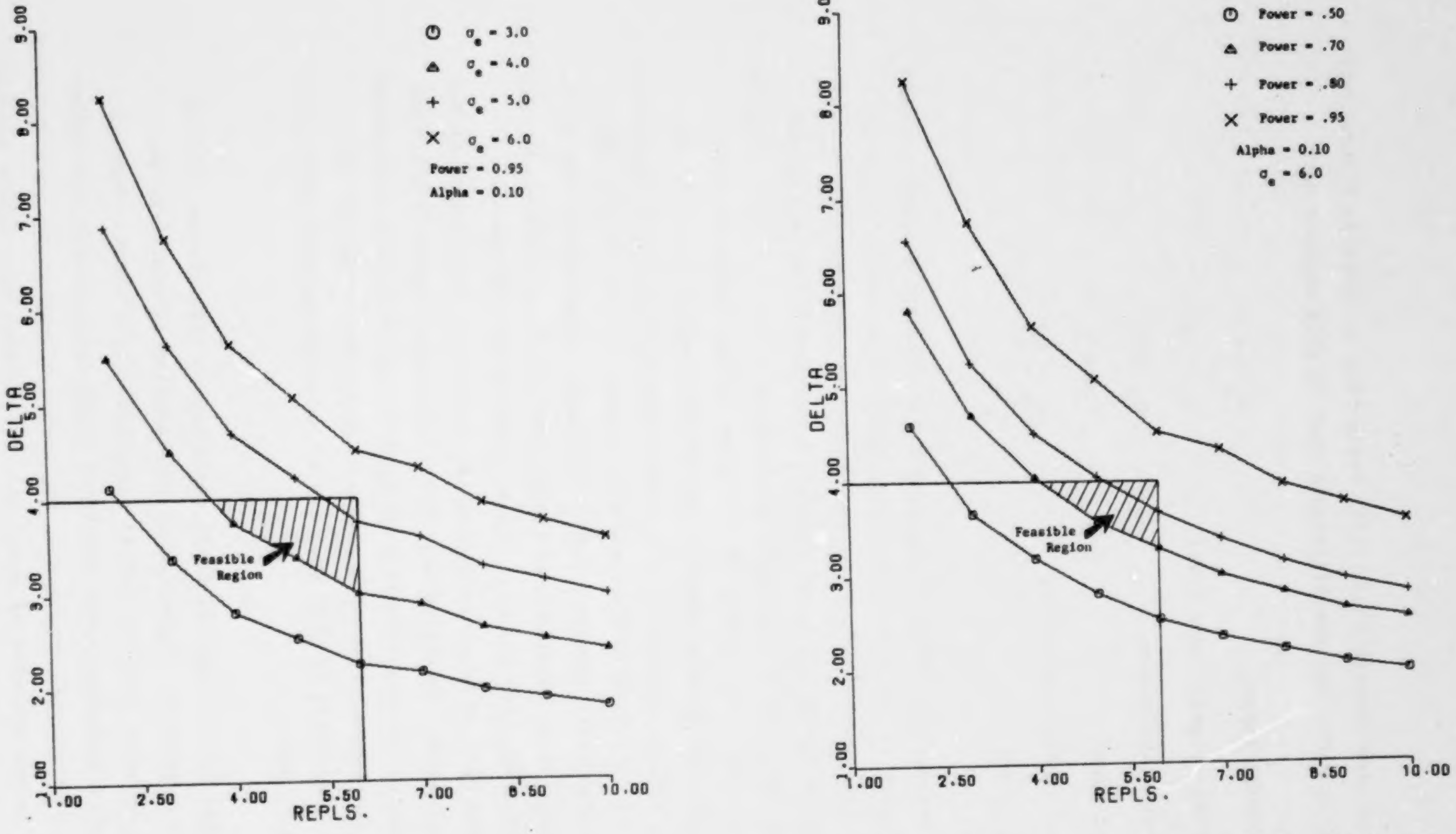
If we presume for the moment that the effort expended in an experiment with a given number of treatments is proportional to the number of samples taken, the acceptable level of effort may be specified in terms of the maximum permissible number of replicates (N_A). In addition, we may say that the risk associated with an experiment is acceptable if the significance level is equal to some value (say .10) and the power for testing a particular main effect impact is greater than some minimum permissible value (say .95). With these definitions we may evaluate the set of all feasible experiments.

A typical experimental feasibility evaluation is shown in Figure 3-10 for a 4x2x2x2 layout. The specifications used in this example are:

∘ Significant impact	$\Delta_s = 4.0$
∘ Acceptable significance level	$\alpha_o = .10$
∘ Acceptable replicates	$N_A = 6$
∘ Nominal power	Power $\geq .70$
∘ Nominal standard deviation	$\sigma_e \geq 4.0$
∘ Number of blocks	$L = 1$

Figure 3-10a defines the range of residual error standard deviation values which give a power of .95 for testing the impact of the first (4 level) factor in the experiment. Since our specifications indicate that σ_e will be greater than 4.0, the region of feasible experiments consists of all layouts having less than 6 replicates which allow a Δ of 4.0 to be detected with a power of .95. This region is indicated in the figure by cross-hatching. Figure 3.10b shows another view of the feasible region. In this case, the region consists of all layouts having less than 6 replicates which allows a Δ of 4.0 to be detected with a power of .70 or more when $\sigma_e = 6.0$. The two plots together give a reasonably good feeling for the marginal sensitivity of experimental feasibility to the specifications imposed by the investigation. For example, Figure 3.10a shows that two more replicates would have to be added to provide acceptable performance if the residual error standard deviation were 6.0. Similarly, Figure 3.10b shows that replacement of the requirement that $\Delta_s = 4.0$ by the milder requirement that $\Delta_s = 5.0$ increases the power obtainable with 6 replicates to .95. Numerous other trade-offs of this sort can be investigated with appropriate power or delta curves.

The concepts and statistical tools described in this chapter provide the means for a systematic evaluation of the sampling alternatives and experimental procedures proposed for an intertidal field study. Some of the simpler design parameters--the specified significance level, the number and allocation of factorial treatments, the number of replicates--are included quite directly in our analysis of power and risk. Other more subtle aspects



a) Feasible Region for Various Values of σ_e

b) Feasible Region for Various Values of Power

FIGURE 3-10

Feasibility Evaluation for a 4x2x2x2 Layout of design--quadrat blocking and location, the use of concomitants, the selection of *which* factors to investigate--are only indirectly included through their influence on the residual error variance. In a practical experimental design, it is generally best to concentrate first on minimizing residual error through blocking, redefinition of factors, use of concomitant variables. or other similar innovations. A reasonable estimate for the resulting residual error (preferably based on experience with a related sequential experiment) should then be used to evaluate feasibility and guide decisions about replication, significance level and other design parameters which are easily controlled.

Although there are many uncertainties involved in this highly judgmental approach to design, we believe there is no realistic alternative. The intertidal environment is too complex to allow us to manufacture simple-minded rules of thumb which apply to all situations. Each experimental program will require a unique approach which accounts for local factors and anomalies that are found nowhere else. Detailed examples which illustrate this point are presented in Chapter 4.

3.4 SUMMARY OF STATISTICAL DESIGN CRITERIA

We have attempted to show in this chapter that statistically valid experiments must first be scientifically valid. The models, blocking schemes, and sampling allocations used in the field inevitably depend on subjective decisions made by the researcher, decisions which rely on experience, ecological understanding, and intuition. We cannot, therefore, ever expect to have a simple set of ground rules or instructions for the design of intertidal experiments. At best, we can outline some general design criteria which serve to remind the investigator of the many factors which must be considered when laying out an experiment. A list of such criteria (or guidelines) is provided in this section, primarily to summarize the many points raised in Sections 3.2 and 3.3. Step-by-step examples of how our design criteria can be applied to a typical intertidal experiment are provided in Chapter 4 and extensive details on specific design questions are given in standard references such as Cox (1958), Cochran and Cox (1957) and Scheffe (1959).

- *The experimenter should lay out a relatively short sequence of separate field studies which each contribute to a clearly defined overall research objective.*

  The sequential approach to experimentation allows the researcher to identify natural processes acting at the experimental site before perturbations are applied. It also provides a chance to refine uncertain experimental models, estimate residual error variance, and test field methodology.

- *The design of each experiment in the experimental series should begin with a careful enumeration of the dependent and independent variables to be considered.*

  It is useful, at the outset, to make a list of all the environmental variables (physical, chemical and biological) which could affect the biological variables of primary interest. These variables form the basis for the experimental procedures and sampling designs which follow.

- *The number of factors, factor levels, and concomitant variables used in the experimental model should be kept as small as possible.*

  It is best to use the same number of levels (usually 2 or 3) for each classified factor unless special considerations require a larger number of particular factors. Although concomitant variables may help to reduce residual error, they should not be added indiscriminately since they can increase experimental cost substantially.

- *The effects of unavoidable nuisance variables should be randomized through the use of pragmatically designed blocking strategies which minimize within block variability and maximize between block variability.*

  It is important to note that experimental units and blocks do not have to be laid out in regular grids or patterns in the interests of "objectivity." Rigid blocking strategies only serve to increase residual errors unnecessarily and do not make experimental conclusions more statistically valid. Although a single blocking arrangement may not deal equally well with all nuisance variables, one based on obvious abundance variations or terrain features will probably be a good overall compromise. Whenever the designer is in doubt about spatially or temporally variable nuisance factors a randomization scheme should be developed to deal with them.

- *The relationships between observations, experimental factors, and potential sources of experimental error should be defined in an experimental model.*

  The experimental model is not just a formality (although it is often treated this way) but is, rather, a reference which allows the designer to deal with unique circumstances which may not be included in standard sampling plans. It helps the designer identify interactions which can be lumped with residual error or confounded to provide better main effect test power. Most importantly, it is the basis for all sums of squares computations used in the analysis of variance and for the degree of freedom and non-centrality parameter computations used in performance evaluations.

- *For intertidal applications, randomized block sampling layouts offer the most flexibility and are easiest to use.*

  Randomized block design procedures are well suited for factorial model studies since they offer the designer considerable flexibility in selecting the number of replicates to be used and the interactions to be emphasized in the analysis of variance. Other designs, such as Latin or lattice squares, can be adapted to factorial applications but the additional computational complexity arising when these designs are confounded or otherwise modified for multi-variable problems is usually not justified.

- *Complete block designs are highly desirable in studies with only a few factors, when the experimental site is relatively uniform, or when interactions may be important. They are, however, usually not the most efficient designs in larger factorial experiments.*

  Complete block designs require a considerable amount of sampling in order to provide information on all factorial interactions. Usually, interactions higher than first or second-order are difficult to interpret even when they are accurately estimated. Often, these interactions are insignificant. In such cases, the information offered by complete block methods may be more than is really needed. Complete block designs can require very large block sizes when more than 4 or 5 factors are included in the experimental model. Since large block sizes are more likely to have significant within-block nuisance variations, they should be avoided whenever possible.

- *Practical limitations on quadrat placement and blocking in the intertidal environment should be considered in the early stages of an experimental design.*

The practical problems associated with executing particular treatment manipulations, particularly those relating to petroleum application, can severely limit the experimenter's ability to lay out ideally randomized blocks of sampling quadrats. Although some incomplete blocking techniques may be able to reduce these problems by sacrificing certain factorial effects or interactions, we feel it is generally better to redefine the objectives and factors of the experiment than to resort to restrictive incomplete or non-randomized methods.

- *Selection of a test decision threshold is an inevitably subjective process which should be carefully considered by the experimenter.*

  The decision threshold selected for testing a particular impact hypothesis determines, to a large extent, whether or not the impact of interest will be judged significant. Ideally, this threshold should be chosen to minimize the normalized decision risk for the hypothesis. Since the decision risk depends on costs which may be difficult to determine, a common practical alternative is to derive the threshold from a specified significance level (type I error). If this method is used the experimenter should attempt to balance type I and type II errors so that the test is not unfairly biased for or against detection of impacts. *It is not sufficient to use a significance level of .05, .10, or any other value simply because this is accepted practice.*

- *Experimental feasibility can be objectively evaluated only if it is defined in terms of specific design parameters.*

  In a qualitative sense, a feasible experiment is one which can detect significant ecological effects (e.g., abundance changes) with acceptable risk performance for an acceptable expenditure of effort. In order for the terms significant and acceptable to be properly defined, the experimenter must specify:

  - The level of effect considered to be significant (Δ_s).
  - The maximum risk, or, equivalently, the power and significance level considered to be acceptable.
  - The maximum number of replicates, treatments, and blocks considered to be acceptable.

  In addition, the experimenter must estimate a reasonable value, or at least a range of values, for the residual error variance to be obtained in the proposed experiment.

To conclude our discussion of the statistical aspects of experimental design, we review the three questions posed at the beginning of the chapter in Section 3.1. We attempt to give brief answers to each question based on the design and performance evaluation concepts presented in Sections 3.2 and 3.3.

- *To what extent can traditional experimental design methods be applied to intertidal and particularly to petroleum-related field studies?*

  The traditional methods of factor selection, randomization, blocking, and performance evaluation discussed in this chapter are directly relevant to intertidal experiments, as illustrated in the various examples presented. The variability, patchiness, and complexity of intertidal species and the difficulties involved in applying petroleum as a treatment require innovative and flexible approaches to factor selection and quadrat placement, but they do not invalidate in any way the basic concepts developed in traditional design theory.

- *What standards must be met for a particular intertidal experiment to be statistically valid?*

  The experimental layout must be properly randomized so that within-block nuisance variations are minimized and the experimental model must explain a reasonably large fraction of observed sample variability (i.e., the residual error variance should be small relative to the measurement variance). In addition, the experimental model, sampling layout, and field methodology should be consistent with the researcher's understanding of ecosystem behavior.

- *How much effort (or cost) must the researcher spend to verify that a particular ecological impact is statisfically significant?*

  This depends on the level of effect the researcher considers significant, on the risk the researcher is willing to accept in his decisions, and on the level of residual error variability present after the best available experimental model has been fit to the data. Systematic techniques are available for relating all of these design specifications to sample size (or cost) for a given experimental layout. The feasibility and desirability of any experiment are determined by the relative values of the scientific benefit which might be gained and the costs which must be expended. When a system is highly complex and poorly understood, the expenditures of time, money and effort required may be large but the potential for gaining important new information may be significant.

4. EXAMPLE DESIGNS OF ROCKY SHORE OIL EXPERIMENTS FOR ZAIKOF BAY, ALASKA

4.1 INTRODUCTION

We now turn to the task of illustrating the use of the design criteria and methods developed in Chapters 2 and 3 in a realistic example. The results yield further understanding of the experimental design problem and demonstrate the role of quantitative design methods in these kinds of problems. In addition to demonstrating the application of ecological and statistical design criteria, we also display alternative designs for a few particular experiments. These displays include an assessment of the tradeoffs between experimental effort and information return. Although our examples are limited in scope and devised to illustrate the role and utility of quantitative design methods, they demonstrate typical designs.

The upper intertidal zone, typically dominated spatially be a complex of barnacles, mussels and algae, is the focus of our examples. Our descriptions and approach are strongly motivated and influenced by work reported by Dayton (1971). We have chosen Zaikof Bay, a site on the southern coast of Alaska studied by Zimmerman et al (1976 and 1977) as a representative rocky shore, suitable for conducting field experiments. Our choice in no way reflects a recommendation one way or another as to the ultimate use of Zaikof Bay for oil-related studies.

Because our purpose is to demonstrate design methods, we have not dealt with some important problems associated with implementing a field experiment. Site selection, logistic considerations, and laboratory procedures are aspects of carrying out an experiment which we have not considered. In particular, as we discussed in Chapter 1, we do not address important political problems which could arise in conducting a "controlled oil spill."

Our examples represent a meaningful compromise between exact complex conditions one would expect to find in the field and simple illustrative cases which are only of academic interest. Many of our assumptions may appear expedient, given our objectives. However, they are exactly the kinds of assumptions that are necessary if one wants to use quantitative design methods. It is often necessary to ignore or lump together certain sources of variability known to affect rocky shore communities. For example, desiccation can play an important role as a cause of physiological stress which effects species abundance distributions in rocky shores. As described in subsequent sections, we do not consider desiccation as a separate factor from wave exposure and tidal elevation. Again, such simplifying assumptions are necessary decisions required in arriving at a design. The use of quantitative design methods accentuates and clarifies these kinds of assumptions and choices.

4.2 EXPERIMENTAL SITE DESCRIPTION - ZAIKOF BAY, ALASKA

In Section 1.5 we described some features of rocky shore habitats and communities. We now consider Zaikof Bay specifically. Our description and understanding of the rocky shore community at Zaikof Bay is based on data reported by Zimmerman and Merrell (1976). We do not have first-hand knowledge of Zaikof Bay and have relied almost entirely on the data as it has been reported to the National Oceanographic Data Center (NODC). As a result some features of the Zaikof Bay site may be overlooked or inaccurately described. Our designs can be readily updated to account for additional knowledge about the site not presently available to us.

Zaikof Bay can be classified as a semi-protected boulder beach with a lot of relief. We choose for our focus of interest the barnacle-mussel zone, typically occuring within the tidal region of 0-8 ft. above mean lower low water (MLLW). Following Dayton's (1971) description we assume that the principal sources of physical disturbance are wave exposure and desiccation. We assume there is no significant log battering and that variations in temperature and salinity do not affect the community.

The principal sessile species occuring at Zaikof Bay in this zone are several algae, two barnacles--Balanus glandulus and B. cariosus--and the mussel Mytilus edulis. The chiton Katherina tunicata, the limpet Collisella pelta, and the urchin Strongylocantratus droebachiensis are the predominant herbivores. Nucella lamellosa, Lepastarias hexactis and Pisaster comprise the important carnivorous predators. Table 4.1 is a list of selected species which constitute the most prevalent taxa at Zaikof Bay. Table 4.2 lists taxonomic groups not included in our selected species lists.

The important biological interactions which are functioning in the community are competition for "primary" space among sessile species and predation by Katherina, Collisella, Nucella and Pisaster. Dayton (1971) describes these interactions in detail for the kind of rocky shore we have in mind. In general, the limpet guild plays a particularly important role because it can efficiently graze settling algae and sporelings, as well as prey on and dislodge immature and young barnacles. Experimental evidence such as Dayton's demonstrates the importance of physical and biological disturbances in rocky shore communities. The interplay of dynamic processes--physical disturbance, biological interaction and settlement/recruitment--result in a dynamic, highly variable system which exhibits significant spatial patchiness, as well as large temporal fluctuations.

For the purposes of our examples we choose the algae-barnacle-limpet interactions as a focal point. We are motivated to make this choice for several reasons. First, we want to work with a sub-system which is observable and controllable. Secondly, as described in Chapter 2, herbivore-algal disturbances are often observed following oil spills. Thirdly, there is a relatively large amount of experimental experience with this system reported in the literature.

4.2.1 Existing Data from Zaikof Bay

In this section we describe and statistically characterize the existing data set for Zaikof Bay. Our primary interest is to obtain estimates of the residual error variance, σ_e^2, which can be used to generate alternative designs for oil-related experiments at Zaikof Bay.

Table 4-1

SELECTED SPECIES LIST FOR ZAIKOF BAY

ALGAE

Chlorophyta

Ulva lactuca

Phaeophyta

Alaria praelonga

Rhodophyta

Pterosiphonia bipinnata
Odonthalia spp.
Halosaccion glandiforme
Rhodymenia palmata

BIVALVES

Mytilus edulis

BROWSING MOLLUSCS

Katherina tunicata
Collisella pelta
Littorina sitkana
Littorina scutulata
Margarites helicinus
Lacuna marmorata

PREDATORY MULLUSCS

Nucella lamellosa
Searlisa dira

BARNACLES

Balanus cariosus
Balanus glandula

AMPHIPOD

Amphithoe rubricatoides

STARFISH

Laptasterias hexactis
Pisaster

SEA URCHINS

Strongylocentrotus droebachiensis

Table 4-2

TAXONOMIC GROUPS NOT INCLUDED IN SELECTED SPECIES LIST FOR ZAIKOF BAY

Bryozoa
Platyhelminthes
Rhynochocoela
Echiuroidea
Nematoda
Sipinculida
Oligochaeta
Arachnida
Insecta
Pycnogonida
Brachiopoda
Urochordata
Holothuriodea
Anthozoa

At the time of this writing, the available intertidal data for Zaikof Bay consists of 15 samples collected on September 12 and 13, 1974. These data are reported in Zimmerman and Merrell (1976) and available on magnetic tape from NODC. Additional samples collected at Zaikof Bay in 1975 are not included in our analysis, because we have not received the necessary tapes from NODC.

The 15 samples comprising the data set (Table 4-3) were obtained by two different sampling methods. Ten samples collected on September 12, 1974 were obtained by the so-called Myren-Pella Method (Zimmerman and Merrell, 1976). This method involves sketching a facsimile of the area to be sampled, and the biotic zonation, on a sheet of mylar plastic. Numbered, homgeneously arrayed dots are then placed on the sketch. A random number table is used to choose the dots which are projected into sampling locations. Numbered arrows are then placed at the corresponding locations on the rock face and sampling with 1/16 m^2 sampling frames follows.

A nested quadrat sampler was used to obtain five samples on September 13, 1974. This frame consists of 16 squares, each 1/64 m^2. Different sized areas are collected (see Table 4-3). The resulting data can be used to determine the variability among sample sizes and assess the adequacy of different sample sizes.

Of the 15 samples available, eight are used in the following statistical analysis. Sample #6 is deleted from the analysis because the quadrat area is unknown. Sample #3 is deleted because there is no replicate sample for the same tidal elevation (1.52m) and variability due to tidal elevation is a principal concern. Finally, we have omitted the nested quadrat data (samples 11 thru 15) from the analysis for two reasons. First, variability with quadrat size has been analyzed by Zimmerman and Merrell (1976) and they demonstrate that a 1/16 in^2 sampling frame is satisfactory. Secondly, the effect of quadrat size is of no particular interest experimentally (beyond choosing an appropriate fixed size) and the available data are insufficient to analyze the interaction between tidal elevation and area because nested quadrat data are available for only a single elevation. Thirdly, the nested quadrat data are collected on a different day, which may introduce additional variability, for which we cannot adequately account with the existing data.

Table 4-3

ZAIKOF BAY DATA SET

September 12 and 13, 1974

SAMPLE NO.	DATE	METHOD	ARROW NO.	QUADRAT AREA (m^2)	TIDAL ELEVATION (m above MLLW)
1	9/12/74	MYREN-PELLA	E10	.0625	1.82
2	↓	↓	W10	.0625	1.82
3			D12	.0625	1.52
4			D15	.0625	1.21
5			D16	.0625	1.21
6			D19	Not reported	.60
7			D22	.0625	.60
8			Z22	.0625	.60
9			D23	.0625	.91
10			Z23	.0625	.91
11	9/13/74	NESTED QUADRAT	Z 1	.0156	1.21
12	↓	↓	Z 2	.0156	1.21
13			Z 3	.0313	1.21
14			Z 4	.0625	1.21
15			Z 5	.1250	1.21

Information taken from magnetic tape provided by NODC

4.2.2 Statistical Analysis of the Zaikof Bay Data

Table 4-4 shows for counts and wet weight, the mean, standard deviation, variance, coefficient of variation (standard deviation/mean) and frequency of occurrence for each selected species computed from the reduced data set of eight samples, treating all eight samples as replicates. Note that the coefficient of variation ranges from 95% to 282% for counts and 19% to 282% for wet weight, demonstrating the relatively large variability in the data. Also, note that most of the species occur in only a few of the eight samples.

The difference in frequency of occurrence between numbers of organisms and wet weight is not explained in available data documentation and is apparently dependent on the characteristics of the species involved. Some algal species are difficult or impossible to measure by counts. For most invertebrates wet weight is a less reliable measure of abundance than counts.

Wet weight measurements often range over four or five orders of magnitude, whereas counts rarely range over more than two orders of magnitude for a given species. For convenience we transform wet weight measurements by log (y+1).

The variances given in Table 4-4 are estimates of residual error variance in the model of measurements, y_i, given by

$$y_i = \mu + \varepsilon_i \qquad i = 1, 2, \ldots 8 \qquad (4-1)$$

where, μ = grand mean of all measurements of a given species

ε_i = residual error with zero mean and variance of σ_e^2

We are interested in finding an alternative model for the measurements which accounts for part of the variability and thereby reduces σ_e^2. A potential source of variability within the reduced data set is tidal elevation. We can hypothesize the following alternative model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \begin{matrix} i = 1, 2, 3, 4 \\ j = 1, 2 \end{matrix} \qquad (4-2)$$

where, α_i = the effect due to tidal elevation i

Table 4-4

STATISTICS OF SELECTED SPECIES AT ZAIKOF BAY

	COUNTS					LOG(WET WEIGHT + 1)				
SPECIES IDENTITY	MEAN	STD DEV	VARIANCE	COEF VAR	FREQ	MEAN	STD DEV	VARIANCE	COEF VAR	FREQ
Ulva lactuca	.25	.53	.28	213.81	2	1.43	1.25	1.56	87.38	5
Alaria praelonga	.00	.00	.00	.00	0	.00	.00	.00	.00	0
Halosaccion glandiforme	2.71	7.38	54.46	271.87	2	1.88	1.69	2.86	89.82	5
Rhodymenia palmata	2.00	3.61	13.03	180.77	3	2.21	1.61	2.60	73.15	6
Pterosiphonia bipinnata	.00	.00	.00	.00	0	.82	.93	.86	113.28	5
Odonthalia dentata	.50	.92	.85	185.16	2	1.57	1.56	2.43	99.41	5
Odonthalia kamtschatica	.00	.00	.00	.00	0	.00	.00	.00	.00	0
Odonthalia washingtoniensis	.00	.00	.00	.00	0	.45	.85	.73	188.28	2
Katherina tunicata	.06	.12	.01	185.16	2	.80	1.50	2.24	186.15	2
Mytilus edulis	31.42	33.61	1129.81	106.98	6	3.26	2.12	4.48	64.93	6
Collisella pelta	5.90	5.66	32.00	95.94	8	2.55	.49	.24	19.14	8
Margarites helicinus	11.08	17.03	290.12	153.78	5	1.00	1.12	1.25	112.17	5
Littorina sitkana	38.84	88.97	7915.66	229.04	5	1.19	1.66	2.76	139.35	5
Littorina scatulata	3.49	7.37	54.34	210.96	3	.64	1.02	1.04	159.73	3
Lacuna marmorata	41.31	99.79	9957.63	241.57	5	.87	.96	.91	110.45	5
Nucella lamellosa	1.90	4.16	17.34	218.77	3	1.31	1.83	3.33	139.74	3
Searlisa dira	.00	.00	.00	.00	0	.00	.00	.00	.00	0
Balanus cariosus	.19	.29	.08	155.33	3	1.05	1.47	2.17	140.02	3
Balanus glandula	5.58	6.95	48.27	124.41	7	2.69	1.26	1.59	46.91	7
Amphithoe rubricatedes	.56	1.31	1.72	233.67	2	.25	.53	.28	209.45	2
Luptasterias hexactis	1.03	1.60	2.57	155.73	6	1.95	1.34	1.79	68.48	6
Pisaster	.09	.19	.03	198.41	2	1.16	2.16	4.67	186.18	2
Strongylocentrotus droebachiensis	.03	.09	.01	282.84	1	.34	.96	.92	282.84	1

Note: All statistics are for 1/64 m^2 area

We can estimate the residual error variance for this model using a one-way analysis of variance (ANOVA). (Note: In a typical application of ANOVA to this kind of problem the analyst would use F-tests to determine the significance of the tidal elevation effect α_i. Since our concern at this point is simply to estimate σ_e^2, we do not carry out a complete ANOVA). Table 4-5 shows the residual error mean sum of squares (variance) for each selected species for model 4-2. Results are shown for both counts and log (y+1) transformation on wet weight. The values of residual error variance in Table 4-5 can be compared with variance estimates in Table 4-4. Differences demonstrate the extent to which the two models (equations 4-1 and 4-2) account for variability in data. Species for which σ_e^2 in Table 4-5 is significantly less than in Table 4-4 have a spatial distribution which is significantly affected by tidal elevation. Notable examples are Rhodymenia, Odonthalia, Mytilus, Collisella, Margaritis, and Balanus glandula. The values of σ_e^2 given in Table 4-4 and 4-5 are essential to the designs generated in section 4.4.

4.3 EXPERIMENTAL OBJECTIVES AND STRATEGY

A number of important questions arise in our consideration of possible oil-related impacts at Zaikof Bay:

- Can the effects of oil on a rocky shore community be differentiated from other independent sources of variability?
- Does the introduction of oil into this system cause a measurable response in patterns of species distribution and abundance?
- What is the minimum concentration of oil at which significant effects can be observed?
- What minimum level of oil effect can be detected with a particular design?

Most of these questions are addressed by adopting the following overall research objective:

> *To study in the upper rocky intertidal community of Zaikof Bay the role of various amounts and compositions of petroleum*

Table 4-5

RESIDUAL ERROR VARIANCE FOR SELECTED SPECIES AT ZAIKOF BAY USING ONE-WAY ANOVA MODEL ON TIDAL ELEVATION

	Residual Error Variance (σ_e^2)	
SPECIES IDENTITY	COUNTS	LOG(wet wt.+1)
Ulva lactuca	.31	.32
Alaria praelonga	.00	.00
Halosaccion glandiforme	55.02	.40
Rhodymenia palmata	.81	.43
Pterosiphonia bipinnata	.00	1.22
Odonthalia dentata	<.01	1.43
Odonthalia kamtschatica	.00	.00
Odonthalia washingtoniensis	.00	.84
Katherina tunicata	.02	2.61
Mytilus edulis	153.66	3.91
Collisella pelta	14.31	.11
Margarites helicinus	36.10	.69
Littorina sitkana	8364.82	1.14
Littorina scatulata	59.33	1.00
Lacuna marmorata	8981.14	.76
Nucella lamellosa	11.32	1.09
Searlisa dira	.00	.00
Balanus cariosus	.11	3.01
Balanus glandula	28.15	1.44
Amphithoe rubricatoides	1.82	.30
Luptasterias hexactis	2.81	.52
Pisaster	.04	5.44
Strongylocentrotus droebachiensis	.01	.92

substances as a disturbance on distribution and abundance of selected species (or taxa) in the following ways:

- *physical/chemical disturbance*
- *direct competition for space*
- *disruption of competition and predation*
- *interaction with natural physical disturbances such as wave exposure and dessication*

Two important criteria developed in Chapter 2 are reflected in this objective. First, we focus our attention on population and community level effects, as exhibited by abundance and distribution patterns of certain species (or other taxa). Secondly, we limit ourselves to the particular sub-system of the rocky shore in Zaikof Bay described in Section 4.2.

Translation of our overall objective into specific experimental designs is our principal task in this Chapter. Guided by the ecological and statistical design criteria developed in Chapters 2 and 3, we follow a process which leads to specific designs intended to help explain the effects of oil spills. It will become evident in what follows that no single experiment is sufficient to accomplish our objective. Rather, a sequential series of related experiments emerge as an experimental approach.

The first step in the design process is to formulate specific experimental hypotheses. To do this, however, we must obtain at least a rough description of community composition (such as provided in Sections 1.5 and 4.2) which help us identify the ecological processes and variables which need to be studied. We then turn to more detailed manipulation experiments designed to investigate hypotheses about the role of natural (as opposed to oil-related) biological and environmental processes in the community. The next step is to establish the feasibility of applying and monitoring oil. Then we proceed to detailed investigations of oil as a treatment, concentrating on comparisons between oiled and unoiled areas. Subsequent experiments can be designed to elucidate the causal mechanisms by which oil affects community structure as well as the relationships between oil and other disturbances.

In Chapter 2 we raised several questions about the feasibility of using oil as a controllable experimental treatment factor. In particular, can oil be applied to units in such a way that actual treatments are known? Does oil applied in one unit contaminate control units or alter the level applied in other nearby units? What is the appropriate scale of oil experiments? What unit size and distance between units should be used? These questions need to be answered as a prerequisite to designing and conducting impact and recovery experiments. To a large extent developing feasible treatment methods and experimental protocols is a trial and error process. The investigator must simply "fiddle around" in the field, testing various experimental methods and devices, developing experience and relying on his or her ingenuity. Statistics and quantitative design methods play a diminished role during this phase of field experiments. In Section 4.4.3 we consider this question further.

It is apparent that results of initial experiments may determine the nature of subsequent investigations. We recommend a cautious, iterative, incremental approach because of the highly complex nature of the system and the tenuousness of existing models of community structure. As more confidence is gained in our understanding of governing phenomena we can move to experiments which are broader in scope and encompass multiple objectives. The design *process* will remain relatively similar.

4.4 DESIGNS

In this section we develop some specific experimental designs intended to meet the experimental objectives outlines in Section 4.3. It is useful to define several discrete experimental phases which would be conducted, more or less, in chronological order:

1. *Preliminary surveys* intended to determine abundance distribution, community composition, and species diversity at the experimental site.
2. *Initial Manipulation Experiments* designed to identify in a qualitative way the factors responsible for observed natural patterns of abundance (in the absence of petroleum).

3. *Oil feasibility experiments or surveys* to establish the feasibility of applying and controlling oil as an experimental treatment.
4. *Oil impact experiments* designed to establish the temporal and spatial variability of selected species when oil is applied as a treatment.
5. *Recovery experiments or surveys* for a long-term followup of the effects of an intentional or accidental spill.

Each stage in this sequence of experiments is, at this time, less well-defined than its predecessors. A myriad of possible lines of investigation could emerge as the results of each survey or manipulation experiment are obtained and analyzed. The examples we present here are intended to illustrate the general process of sequential design. More detailed designs can be developed only as field data actually become available.

The process we follow for choosing the design variables follows from Chapters 2 and 3. First, we develop an explicit qualitative statement of the experimental objectives. From our knowledge and assumptions about possible causal factors in rocky shores, we can select the experimental factors to be included in the experiment. Different levels for each factor define the experimental treatments and establish the range of validity for the results. Some treatments may be represented by natural gradients, such as different tidal elevations. Other treatments arise from specific manipulations, such as predator exclusion devices. Choice of experimental units and observations follow directly from the selected treatments.

After specifying the experimental treatments and units, we can consider alternative layouts for the allocation of treatments to units. For each layout of interest, we use power analysis to evaluate the relationship between probability of obtaining a significant effect and sampling effort. Finally, we can compare alternative designs based on the expected information return (as defined by the model and power analysis) and the total sampling effort.

4.4.1 Preliminary Surveys

In the case of Zaikof Bay, we are fortunate to have some information on community composition and species abundance reported by Zimmerman and

Merrell (1976). This information, summarized in Section 4.2, is precisely the kind we would seek in a preliminary survey of an experimental site. It is useful here to review some of the general sampling principles for such surveys which were, to some extent, used by Zimmerman and Merrell.

As pointed out in Chapter 3, survey studies are concerned with estimation of the mean of a randomly dispersed population. In our example, the populations of interest are species found at Zaikof Bay and are characterized by species counts, wet weight or other abundance-related measures. The quality of the mean estimate obtained in any particular survey depends inversely on the quadrat-to-quadrat variability of the observation. The best single way to reduce this variability is to stratify the sampling procedure--i.e., to group the quadrats into distinct strata which differ with respect to some important environmental variable such as tidal elevation (see the discussion accompanying Figure 3-1). An inspection of the Zimmerman and Merrell (1976) data and the subsequent analysis performed in Section 4.2.2 shows that tidal elevation may be used to stratify the samples in their survey.

In general, sample strata should be selected to account for obvious sources of variability at the survey site--topographic irregularities, differences in dessication or exposure, substrate variations, etc. This procedure provides a useful prelude to controlled experimentation (since the strata may later become experimental blocks) as well as a better estimate of the true abundance and distribution of each species surveyed. With preliminary survey information in hand, the experimenter can proceed with greater confidence to define experimental objectives, factors, treatments, and layouts in the subsequent stages of a sequential research program.

4.4.2 Initial Manipulation Experiments

Initial manipulation experiments are intended to determine or confirm natural patterns of species abundance and distribution and their causes. Even in cases such as rocky shores where previous investigations have led to specific conclusions about causes of observed community structure, Cox

(1958) points out the need to demonstrate agreement with previously established results. In our case, we follow the work of many investigators, especially Dayton (1971) and Paine (1969; 1974). So extensive initial experiments will not be required. Rather, we seek to design some relatively simple experiments which provide a linkage between their reported results and our ultimate goal of investigating effects of oil. This kind of experiment is not a necessary pre-requisite to conducting oil impact experiments. However, we believe it is wise to establish some understanding of the important community processes governing patterns of species abundance and distribution at the site of oil experiments. In addition, because the design of this experiment does not depend on the feasibility of using oil as a treatment factor, it provides our most complete example of experimental design.

Experimental Objective

Results reported by Dayton (1971) and others demonstrate that the patterns of abundance and distribution of sessile species in the barnacle-mussel zone of rocky intertidal habitats is governed by certain physical disturbances and the dominant competitors and predators in the community. In order to confirm this hypothesis for the particular areas in which our oil experiments are conducted, we focus attention on competition among selected algal and barnacle species (Table 4-1) and the role of tidal elevation, Katherina tunicata, Collisella spp., Nucella spp. and Searlisa dira in determining the recruitment and survival of these sessile species.

We can state this objective in terms of specific hypotheses to be tested.

> *There is no significant change in the abundance of selected algal and barnacle species occurring in experimental units due to the following factors and/or their interactions:*
>
> - *differences in tidal elevation*
> - *exclusion or not of Katherina tunicata*
> - *exclusion or not of limpets (Collisella)*
> - *exclusion or not of carnivorous gastropods (Nucella or Searlisa dira)*

Implied in this hypothesis are several constraints on the experiment. We assume that starfish are naturally or experimentally excluded. Only one site is chosen for the investigation. We are not including different wave exposures as an experimental factor. All experiments are begun in Spring of the year and concluded in the subsequent Fall. Possible differences due to different starting dates (month or season) or different years are not investigated in these experiments.

Factors and Treatments

Treatments for this experiment consist of different combinations of the levels of the four factors: tidal elevation, Katherina, limpet and carniverous gastropod abundance. The number of levels of tidal elevation depends on the particular zonation of the study site. Each level should consist of a region in which the spatial distribution of organisms is relatively homogeneous. At Zaikof Bay data exists for four different tidal elevations, so we assume that these four levels adequately represent the site:

Tidal Elevation Level	Height Above MLLW (meters)
1	.6
2	.9
3	1.21
4	1.82

Ideally, the experiment would include several levels of species abundance ranging from complete absence to normal abundance. However, experimentally maintaining abundances at levels between these two extremes is impractical, if not impossible. Therefore, we limit the abundance to two levels--excluded and not excluded.

Alternative Layouts

In the jargon of experimental design, the experiment we are describing is called 4x2x2x2 (or $4x2^3$) factorial experiment. Several possible layouts come to mind. One is a complete block factorial design with blocks of 32 units; another is a confounded factorial design with blocks of 8 or 16 units. A third alternative is to consider each of the four tidal elevations independently and lay out four 2^3 experiments, one at each elevation; or to use a split-plot design, separating blocks into plots according to tidal elevation.

We can rule out this latter alternative because we are specifically interested in main and interaction tidal elevation effects. These are excluded if we use four 2^3 experiments or a split plot layout. Confounded designs are suitable for cases in which we can expect experimental error to be reduced by using blocks which are more homogeneous, or which can be subjected to more uniform techniques, than a complete block. These conditions are not likely to be encountered in this case. The principal expected source of heterogeneity is tidal elevation, which is already included as a factor. At any given tidal elevation variability over clusters of eight 1/16 m^2 units, is unlikely to be reduced by using blocks of two or four units. In experiments where a larger total area is encompassed by the units, confounded layouts may be of more interest.

We are left then with 4x2x2x2 complete block factorial design in which each block contains 32 units. Each unit receiving a different treatment, corresponding to each of the 32 possible factor-level combinations. Allocation of treatments to the units in any block should be random. If replicate blocks are used, these should be positioned randomly (see Chapter 3). Note that this layout allows us to estimate all main and interaction effects for the four factors.

Number of Replicates

Power analysis allows us to determine the number of replicates necessary to have a given probability of obtaining a significant result. As described in Chapter 3, the computation of power depends on the particular statistical test to be performed. In the case of a factorial experiment we use analysis of variance and F-tests to test hypotheses regarding various effects. Given the 4x2x2x2 experimental layout, we determine the degrees of freedom and noncentrality parameter and then calculate F-test power for the experiment.

In general, for a complete block factorial layout, power is largest for main effects and smallest for the highest-order interaction effect. In the present layout, power of F-tests to test for the presence/absence effect of Katherina, Collisella and Nucella are interchangeable because they have the same number of degrees of freedom and the same noncentrality parameter. Therefore, in what follows we use the term power of F-tests to detect a "predation effect" (symbolized by H_p) to refer to the power of the F-test for any one of these three main effects. Because the factor tidal elevation occurs at four levels in this layout, the power of F-tests for tidal elevation effect is lower than for the main effects of predators. In Table 4-6 we compare the power of F-tests as a function of the number of replicates and $\Delta/\sigma_e=1$ for detecting a tidal elevation effect and for detecting a predator effect. In the remainder of this section we use only the power of the F-test for detecting a predator effect, because these are the effects of primary interest. For any given design, power for the main effect of tidal elevation or interaction effects is lower than for the main effect of one of the predator species.

An important parameter in determining the power (and potential success) of any experiment and statistical test is the ratio $\frac{\Delta}{\sigma_e}$. As discussed in Chapter 3, we need an estimate of σ_e in order to evaluate the absolute power of a given experiment. However, parameterizing power on Δ/σ_e provides us with a useful relative measure of experimental performance. Figure 4-1 is a graph of the power of the F-test for H_p versus $\frac{\Delta}{\sigma_e}$ for different replication sizes. It is readily apparent from this figure that for 4 or more replications, the probability of detecting a significant effect is relatively high for a wide range of Δ/σ_e values.

Given an estimate of σ_e^2, the variance of the residual error, we can compute estimates of the minimum detectable effect, Δ, for given replication sizes and powers. Recall that the observations are measures of abundances of different species of algae, barnacles, limpets, etc. We estimate σ_e for a set of selected species in Section 4.2.1 (Table 4-5). The estimates of σ_e^2 shown in Table 4-5 are based on a one factor ANOVA model. This model does not include the effects due to predators, which underlie the 4x2x2x2 complete block layout of the experiment. In using the Table 4-5 estimates of σ_e^2 we must recognize that they are estimates of σ_e^2 for the

Table 4-6

POWER OF F-TEST ON MAIN EFFECTS FOR 4x2x2x2 EXPERIMENTAL LAYOUT IN THE CASE OF $\Delta/\sigma_e = 1, \alpha = .05$

	POWER	
REPLICATES	TIDAL ELEV. EFFECT	PREDATOR EFFECT
2	.59	.98
4	.93	1.0
6	.99	1.0
8	1.0	1.0
10	1.0	1.0

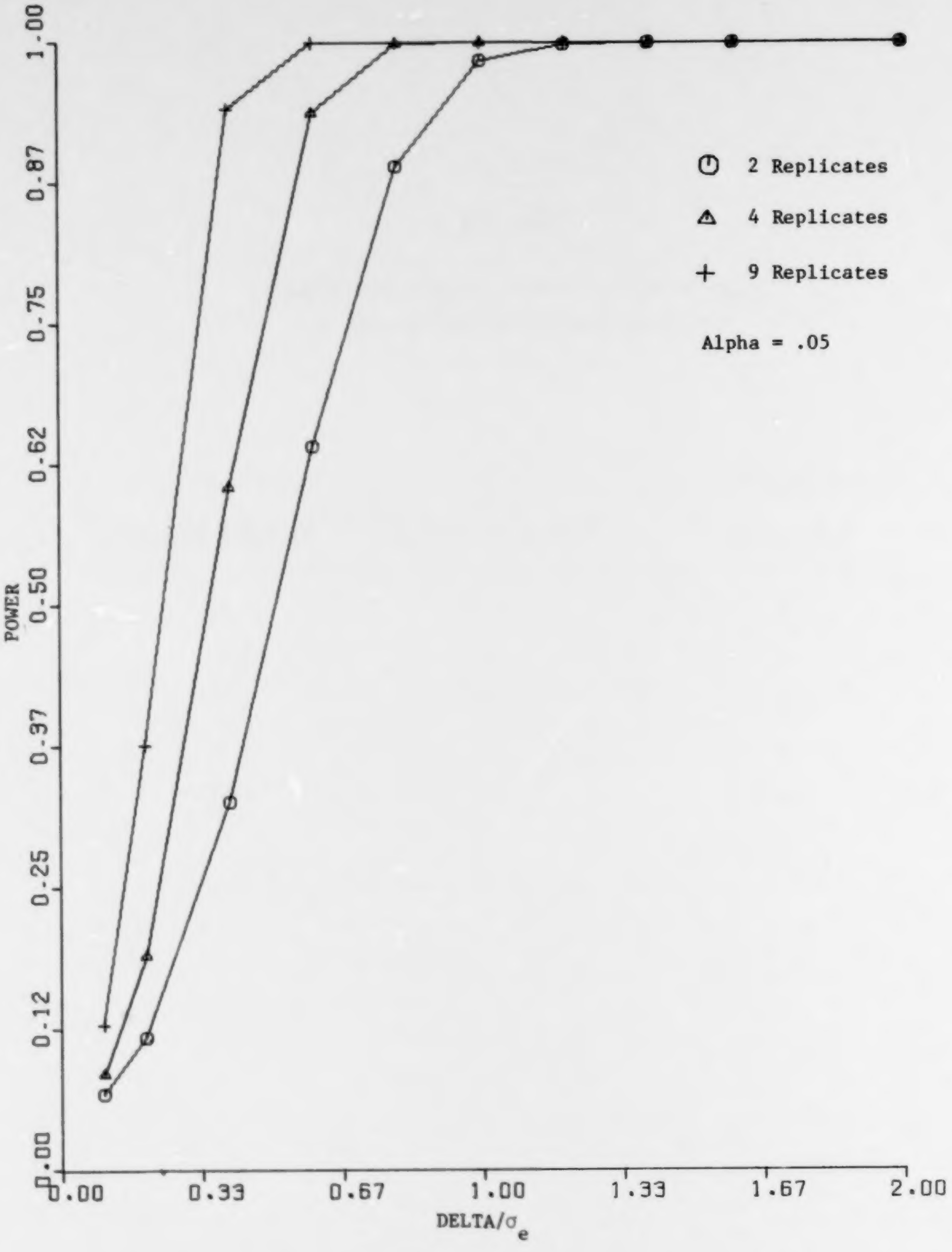


FIGURE 4-1

F-test Power vs. Normalized "predator effect" (Δ/σ_e)

with Different Numbers of Replicates

in a 4 x 2^3 Complete Block Layout

experiment and that data from the experiment itself may yield significantly different estimates. However, these are the best estimates of σ_e^2 based on available data.

For convenience we display minimum detectable change (Δ) as a percentage of the mean value for a particular species. For example, for Balanus glandula our estimate of the mean in Table 4-4 is 5.5 organisms per 1/64 m^2 and of σ_e in Table 4-5 is 5.3 organisms per 1/64 m^2. A change in B. glandula abundance of 1 organism per 1/64 m^2 is an 18% change from the mean. In Figure 4-2a we plot minimum detectable change in B. glandula abundance as percentage of the mean versus number of replicates for powers of .7, .85 and .95 and alpha of .05. Figures 4-2b through 4-2e display similar plots for four other species: Odonthalia, Rhodymenia, Mytilus and Nucella. Note that for Odonthalia, the measure of abundance is log(wet weight + 1). For all other species abundance is measured in counts. Except for Nucella, the detectable % of mean for given number of replicates and power is within a relatively narrow range. This fortuitous result follows from the fact that the ratio of σ_e/mean for each of the species is within a relatively narrow range (.4 - .9).

From Figures 4-2a through 4-2e we can obtain some sense of the economic feasibility of the experiment we have designed. To illustrate, we assume that costs are proportional to the number of replicates and, for example, that we have a budget constraint which limits the experiment to a maximum of 6 replications (192 quadrats). We also assume that we want an .85 probability of detecting a 50% change in the mean of B. glandula at a .05 significance level. The cross-hatched area in Figure 4-2a defines the region of feasible experiments to meet these constraints. As shown in the figure, at least 4 replicates of the 4x2x2x2 experiment are required to provide a .85 probability of detecting a 50% change in the mean of B. glandula, if such changes exist, at .05 significance level. Given a sufficient budget to include 6 replicates in the experiment, a decision maker may opt for the increased power (.95) of detecting a 50% change, which is obtained with 6 replicates. Or, from another point-of-view, 6 replicates allows smaller changes (~40%) to be detected with a probability of .85.

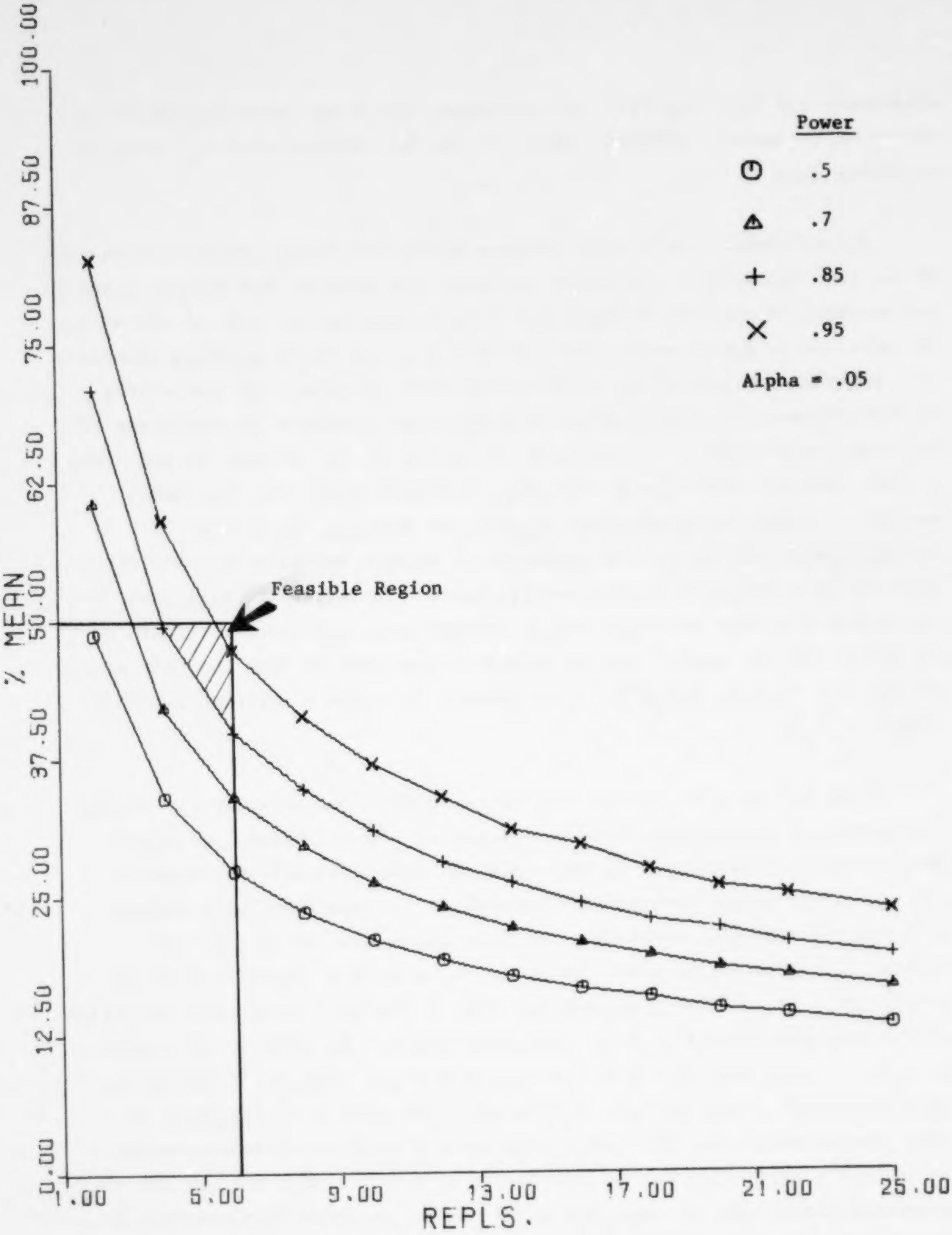


FIGURE 4-2a

Detectable Change as a Percentage of the Mean for

B. glandula vs. Number of Replicates for Different Powers

in a 4 x 2^3 Complete Block Layout

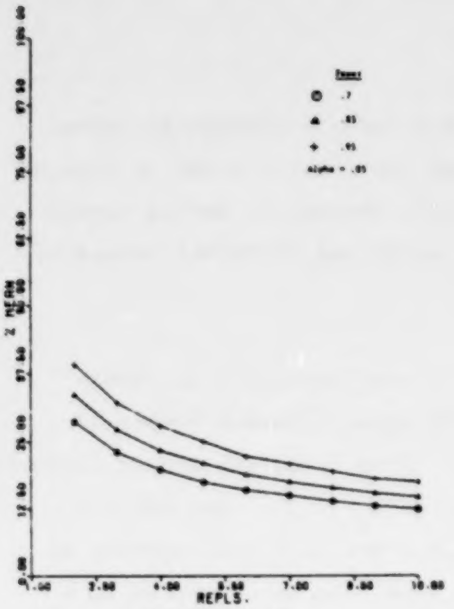


FIGURE 4-2b

Detectable Change as a Percentage of the Mean for Rhodymenia palmata vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout

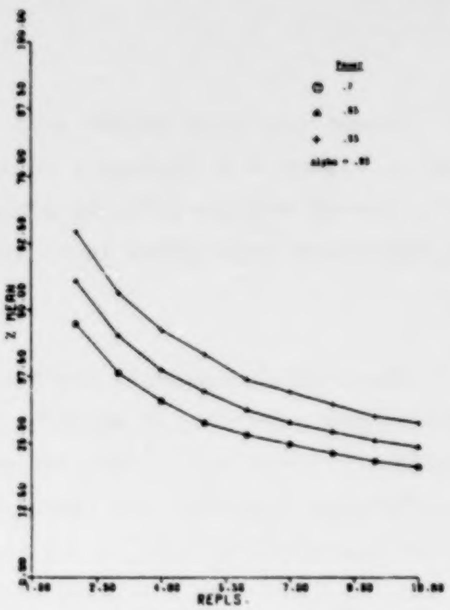


FIGURE 4-2c

Detectable Change as a Percentage of the Mean for Odonthalis dentata vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout

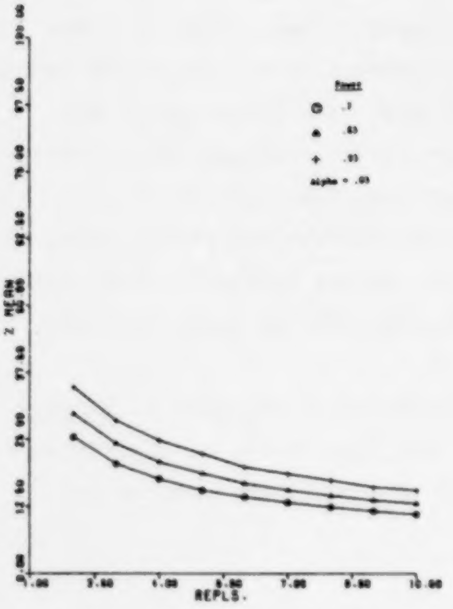


FIGURE 4-2d

Detectable Change as a Percentage of the Mean for Mytilus edulis vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout

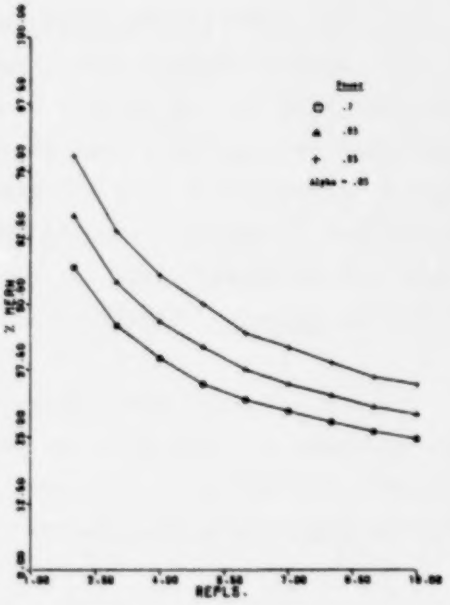


FIGURE 4-2e

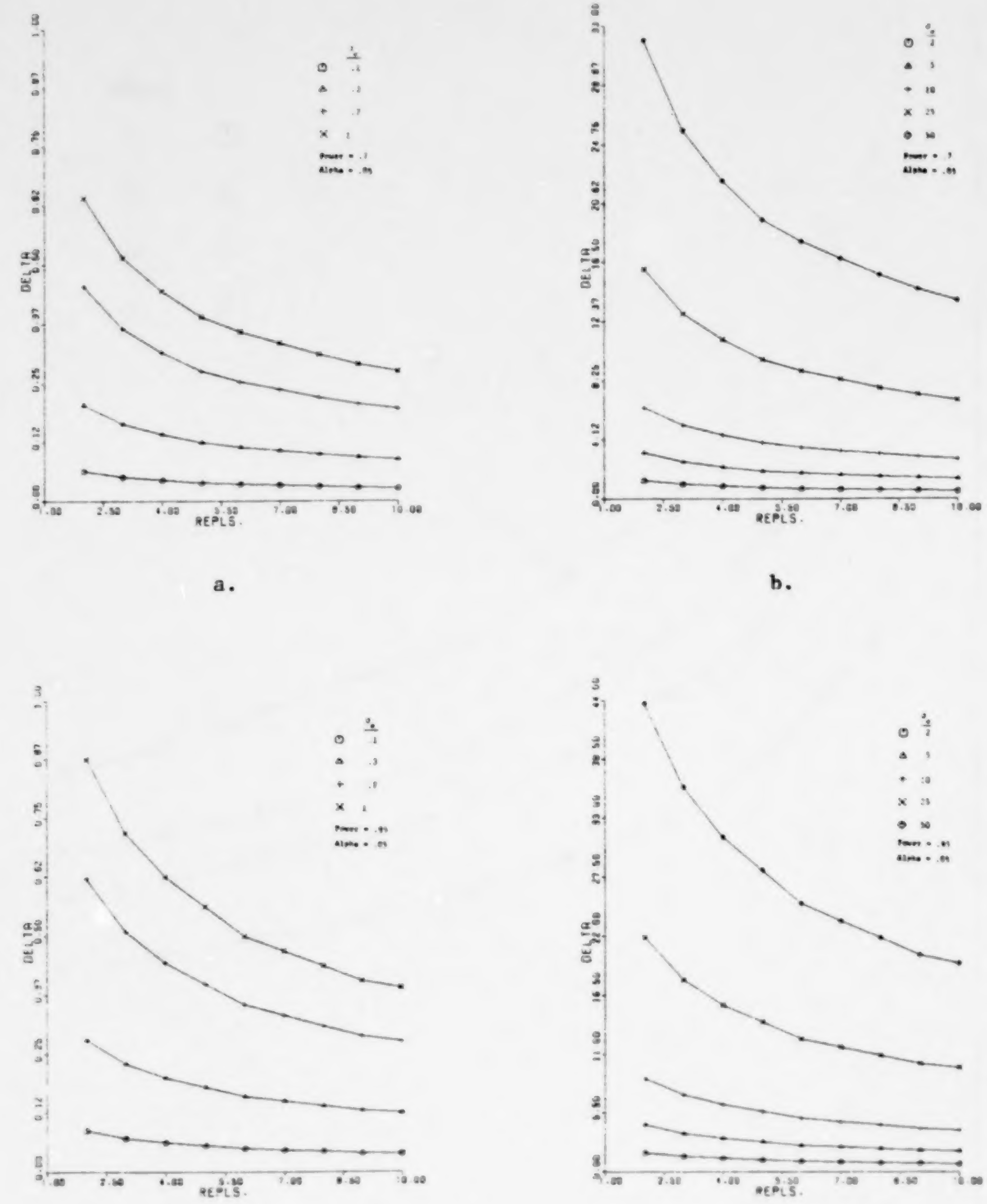
Detectable Change as a Percentage of the Mean for Nucella lamellosa vs. Number of Replicates for Different Powers in a 4 x 2^3 Complete Block Layout

Rather than plot graphs such as Figure 4-2 for all species at Zaikof Bay, in Figure 4-3a through 4-3d we generalize the results shown in Figures 4-2a through 4-2e by plotting minimum detectable change, Δ, versus number of replicates for a given level of power and alpha and different values of σ_e.

The following example demonstrates how to use Figure 4-3 to obtain results for a particular species. Figure 4-3b shows Δ versus number of replicates for $2. \leq \sigma_e \leq 50.$ and power of .7. From Table 4-5 choose a species (say Balanus glandula) and obtain an estimate of σ_e (5.3). Now use the curve corresponding to $\sigma_e = 5.0$ in Figure 4-3b. For different numbers of replicates (say, 6.) the value of Δ obtained (1.8) can be converted to a percent of mean for the given species by using the mean value (5.6) given in Table 4-4. The result of 32% equals that given in Figure 4-2a for the same parameter values. Figure 4-3a shows Δ vs. replicates for $.1 \leq \sigma_e \leq 1.$ and power of .7. Figures 4-3c and 4-3d are for a power of .95. Note that the scaling of the graphs leads to the appearance that Δ is independent of replicates for most values of σ_e.

To this point in the analysis we have assumed a fixed value of alpha of .05. Because alpha is the probability of making a type I error, it represents one kind of risk which the experimenter must take in designing and conducting an experiment (see Chapter 3). The risk represented by a chosen level of significance must be balanced against the risk represented by probability of making a type II error, as established by the chosen level of power. We investigate the role of alpha in the design relative to detecting changes in Balanus. Similar results could be computed for other species.

Figure 4-4 shows detectable change as a percent of the mean of Balanus versus number of replicates for power of .85 and four levels of significance: .01, .05, .10 and .25. The curve for $\alpha = .05$ is exactly the same as the curve in Figure 4-2a for power = .85.



a.

b.

c.

d.

FIGURE 4-3

Delta vs. Replicates for Different Values of σ_e in 4 x 2^3 Complete Block Experimental Layout

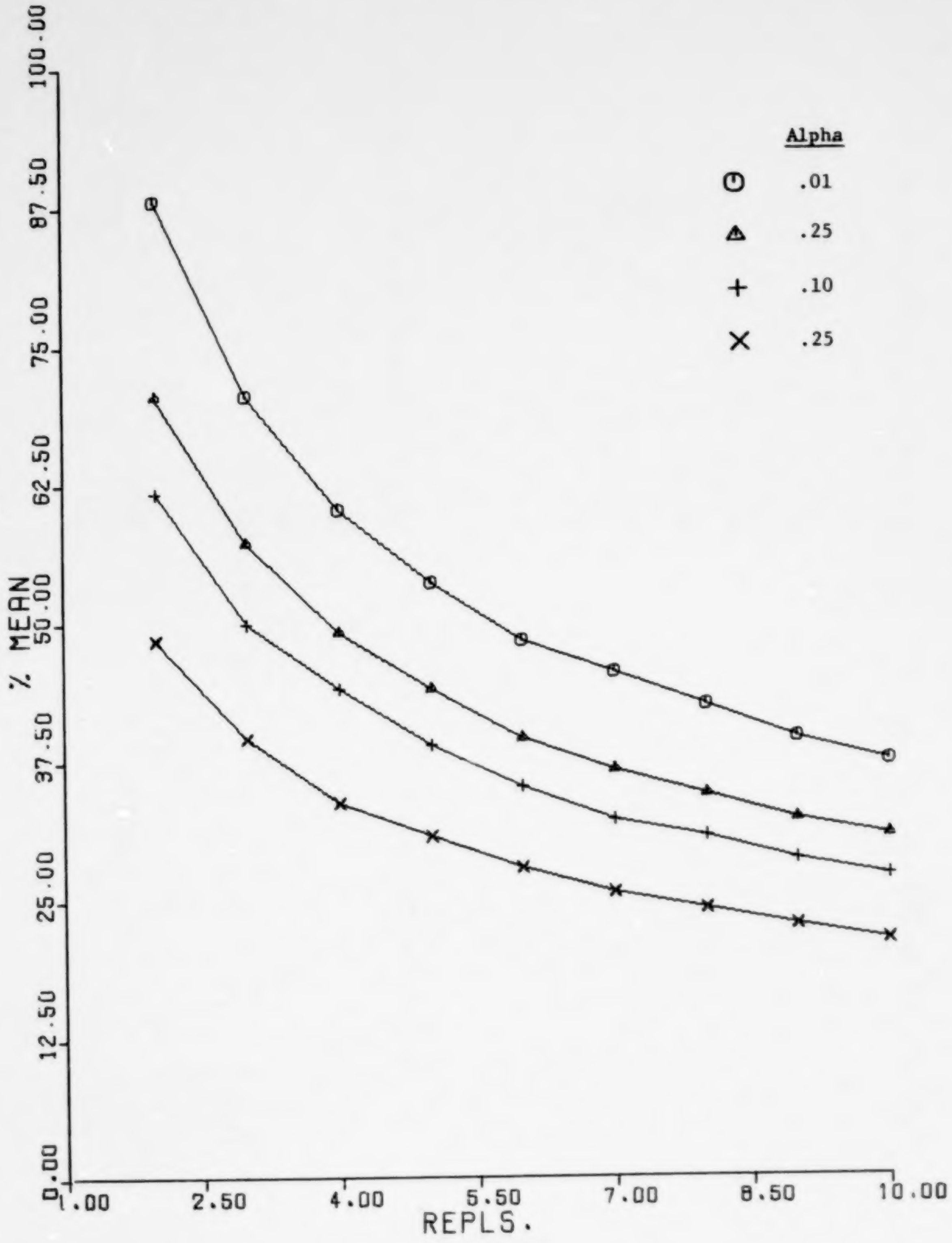


FIGURE 4-4

Detectable Change in B. glandula as a Percentage of the Mean vs. Replicates of a 4 x 2^3 Complete Block Layout for Power of .85 and Various Levels of Significance (α)

Figure 4-5 displays power versus alpha directly for detecting different levels of change in Balanus using four replications. It is readily evident that with four replications relatively small changes in Balanus abundance cannot be reliably detected even at large values of alpha (i.e., a high risk of making a Type I error). To obtain a relatively high power and low level of significance with four replications, the investigator must be willing to accept a large minimum level of detectable change.

As with Figures 4-2a through 4-2e we generalize the results shown in Figure 4-5 by parameterizing the power vs alpha curves on Δ/σ_e. The results are shown in Figures 4-6a and 4-6b for four replicates and 20 replicates respectively. Using means and standard deviations from Tables 4-4 and 4-5 we can convert the generalized Δ/σ_e curves to % of mean curves for a particular species. For example, consider again Balanus glandula with a mean of 5.6 and $\sigma_e = 5.3$. A change in B. glandula abundance equal to .53 corresponds to $\Delta/\sigma_e = .1$ and is equivalent to 9.5% of the mean. Therefore, in Figure 4-6a and 4-6b the curves for $\Delta/\sigma_e = .1$ correspond closely to the 10% of the mean curve in Figure 4-5.

For illustrative purposes we have used B. glandula, one of the less variable species in Table 4-4. Equivalent analysis for more variable or less dense species is also possible. In fact, the interested reader can explore designs for each of the species in Tables 4-4 and 4-5 using Figures 4-1, 4-3 and 4-6 in conjunction with the means and σ_e values given in the tables (or obtained elsewhere for these or any other species).

Figures 4-6a and 4-6b provide further insight to economic feasibility of a given experiment. Assume that budget constraints limit the maximum experiment size to 20 replications of the 4x2x2x2 layout. Then from Figure 4-6b we can evaluate the risk trade-off between power and alpha for detecting certain levels of change. For example, assume we require detecting changes in mean abundance as small as 10% of σ_e ($\Delta/\sigma_e = .1$). Then, the acceptable trade-offs between power and alpha are displayed by the curve for $\Delta/\sigma_e = .1$ in Figure 4-6b. For this particular case it is apparent that a high risk of making a type I error (high α) must be accepted to obtain a relatively high probability of detecting changes equal to 10% of σ_e. For further insight to the risk trade-off between type I and type II errors, see section 3.3.3.

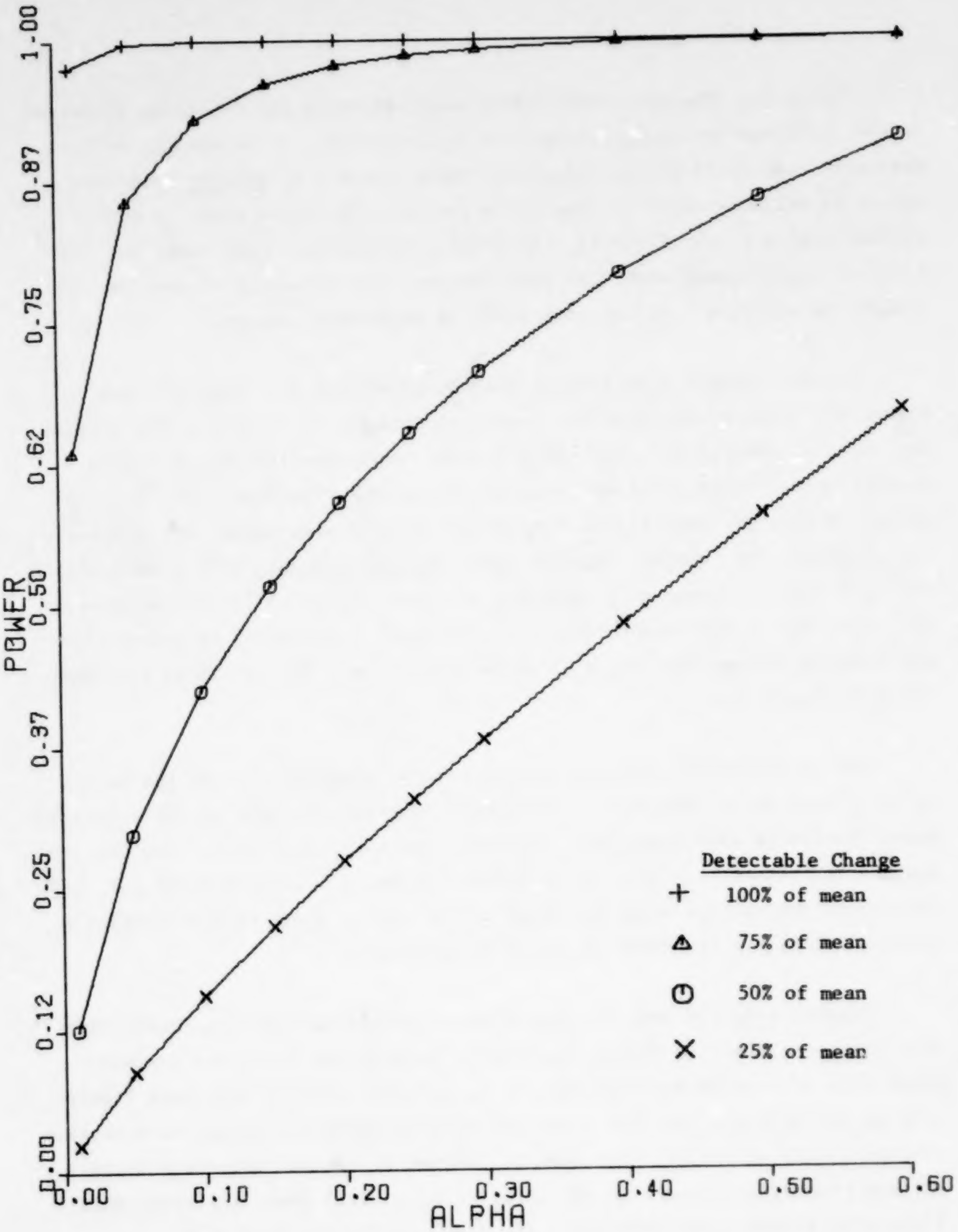


FIGURE 4-5

Power vs. Alpha for Detecting Different Levels of Change in *B. glandula* With Four Replicates of a 4 x 2^3 Complete Block Layout

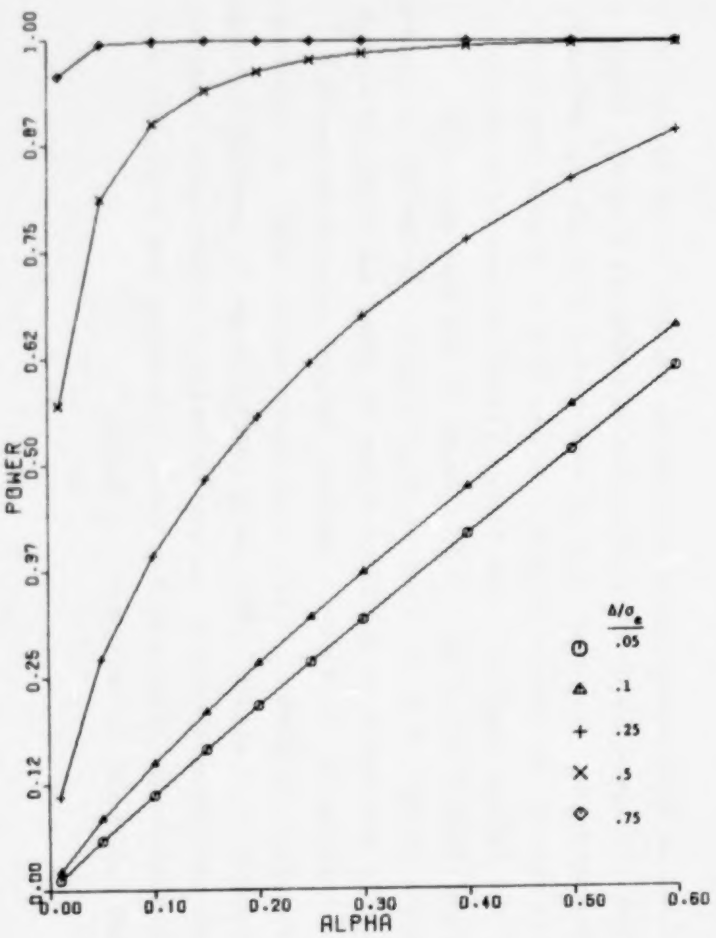


FIGURE 4-6a

Power vs. Alpha for Different Values of Δ/σ_e with Four Replicates of 4×2^3 Complete Block Layout

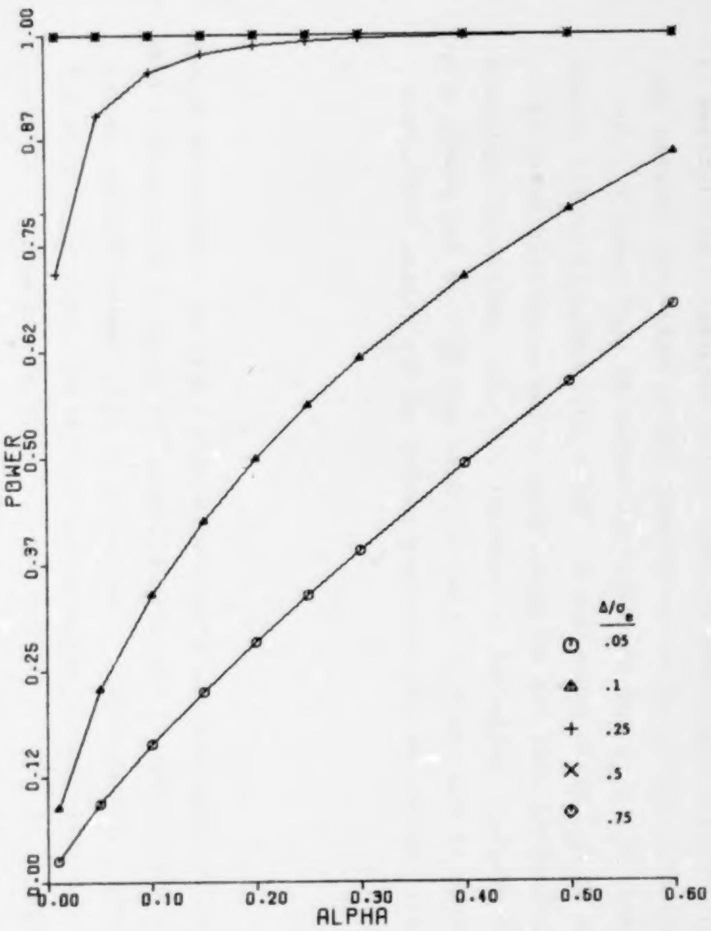


FIGURE 4-6b

Power vs. Alpha for Different Values of Δ/σ_e with Twenty Replicates of 4×2^3 Complete Block Layout

Finally, for this experiment we look at the marginal sensitivities of σ_e to replications. In Figure 4-7 we present the relationship between the maximum tolerable σ_e, σ_{ea}, and the required number of replicates K_a, for various values of Δ_s, given power and α. The relationship is nearly linear throughout, indicating that the marginal gain in the acceptable level of residual error standard deviation is constant for each additional replicate. Given an estimate of the residual error variance and choices for power, α and Δ_s we can readily determine the necessary number of replicates from plots such as Figure 4-7.

Summary

The foregoing analysis has displayed a wide range of alternative design choices within the context of the single layout alternative of a 4x2x2x2 complete block factorial experiment. The choice of this layout follows rather directly from the statement of experimental objectives given in Section 4.3 and our translation of this objective into an experimental model. Alternative layouts--incomplete block, split-plot, confounded designs--do not provide any apparent advantages in this case.

Our ability to apply quantitative statistical methods to the selection of the number of replicates leads to a multitude of trade-off displays among various design variables. The shear mass of computation tends to overemphasize this particular part of the design problem. In any case it appears that 4-8 replicates of the 4x2x2x2 complete block layout provide a reasonable chance (.7 - .95 probability) of detecting 25-75% changes in the mean values of important sessile species at a .05 level of significance. In theory any desired level of change is detectable at specified values of power and alpha, if enough replicates are included in the design. However, budget constraints restrict the size of any given experiment. Given constraints on the number of replicates it may not be possible to detect sufficiently small changes in species abundance with sufficient confidence (power) to warrant undertaking a particular investigation. Particular design choices within these constraints and trade-offs must be made by each individual investigator *and* funder.

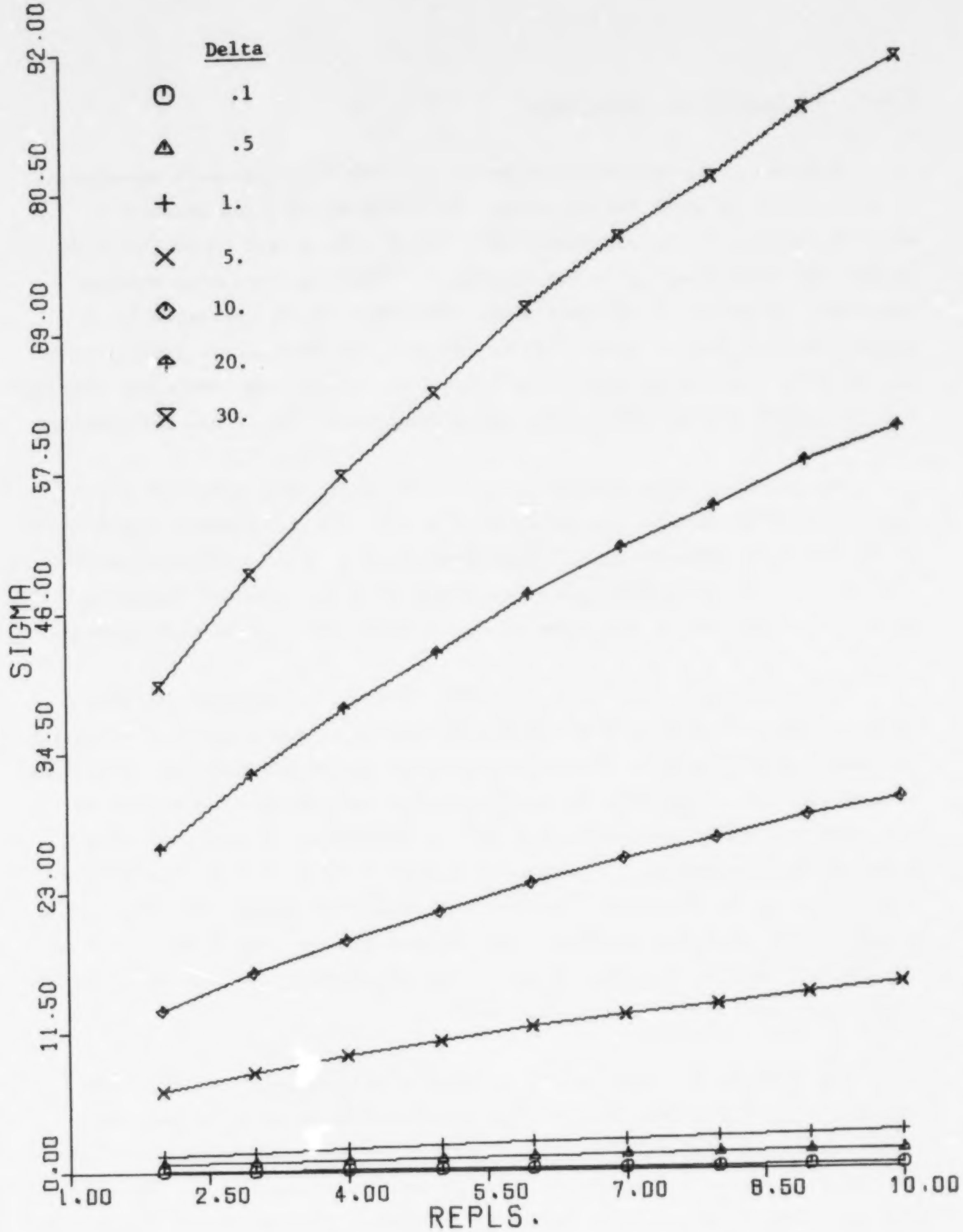


FIGURE 4-7

Marginal Sensitivity of Residual Error Standard Deviation to Replicates for Power .85 and Different Levels of Detectable Change in a 4 x 2^3 Complete Block Experimental Layout

4.4.3 Oil Feasibility Experiments

A prerequisite to conducting effective impact and recovery experiments is the ability to apply oil treatments in a controllable and measurable way. Relatively little experience with oil as a treatment factor in field studies has been reported in the literature. The most extensive studies comparable to the kinds of experiments with which we are concerned in this report are described by Crapp (1971). However, it is apparent from Crapp's descriptions that relatively little control of the oil treatments was obtained and the extent of carryover of oil among experimental units was not measured.

The principal difficulties associated with oil as a treatment factor are discussed in Section 2.4 and Section 3.3.2. For our present purposes we differentiate between difficulties associated with the complex physicochemistry of oil and difficulties associated with applying and confining known quantities and compositions of oil to specified experimental quadrats.

Increasing experience with petroleum chemistry in natural environments suggests to us that with sufficient experience and analytical chemistry hardware feasible methods of sample collection and measurement can be devised by competent investigators. We recognize that considerable "fiddling" in the field may be required and point out the importance of this preliminary phase of experimentation. A necessary outcome of this part of feasibility experiments is the ability to accurately estimate the amount and composition of oil present within a specified experimental quadrat. We do not offer any experimental designs for this aspect of the field study since it is by nature highly empirical and situation-dependent.

The problem of confining oil treatments to particular units raises the most difficult doubts for us about the feasibility of oil experiments. Conceivably, experimental devices and methods can be invented which reliably confine oil to specific quadrats, although it is difficult to say how effective any given device will be before it is actually tested in the field. The alternative which we consider here is to devise experimental designs which account for the "smearing" of oil by tides and other forces over an area outside of its initial application. Such designs (layouts) must ensure the independence of treatments and prevent contamination of controls. The problem

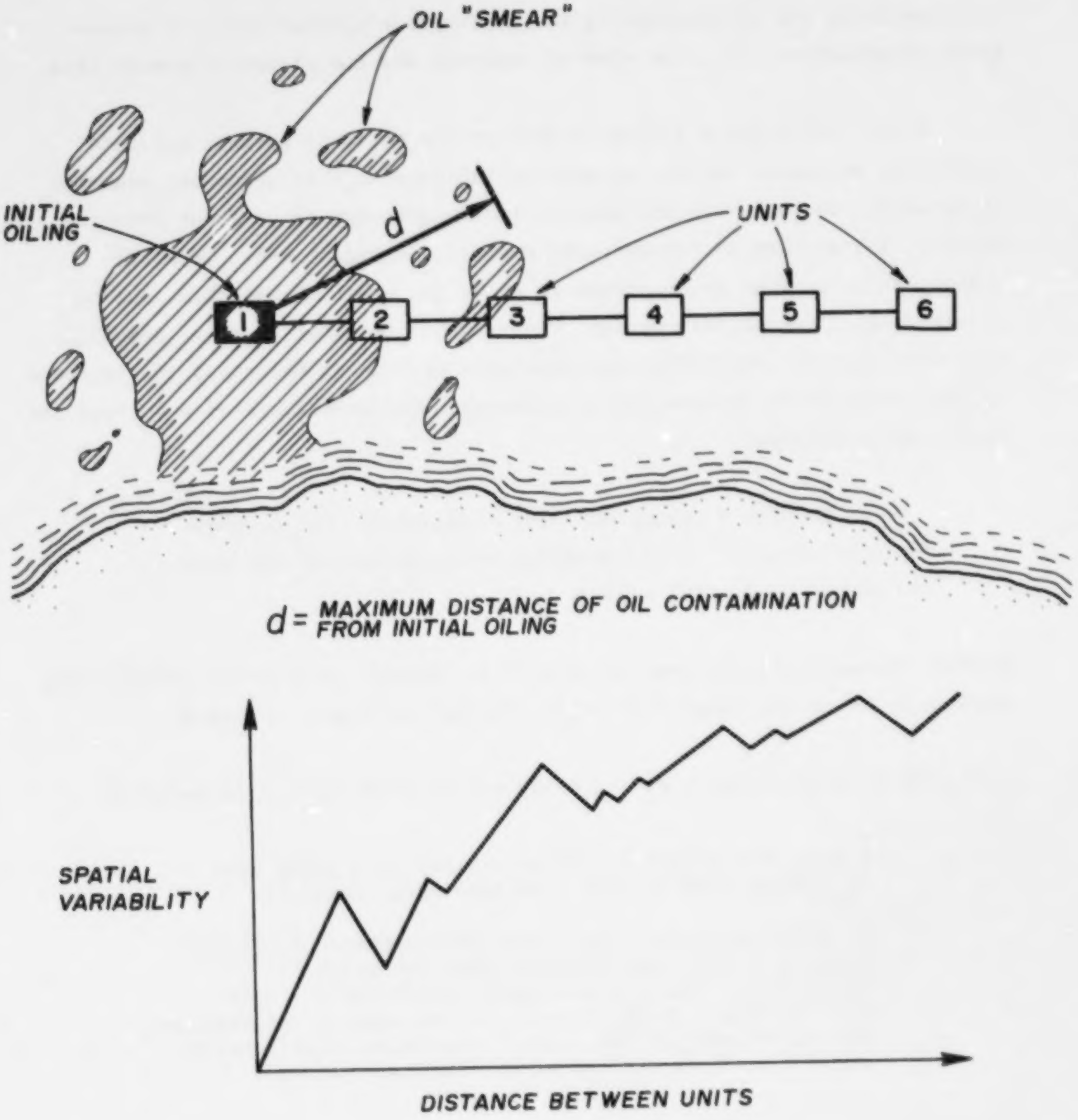

OIL "SMEAR"
INITIAL OILING
d
UNITS
1
2
3
4
5
6
d = MAXIMUM DISTANCE OF OIL CONTAMINATION FROM INITIAL OILING
SPATIAL VARIABILITY
DISTANCE BETWEEN UNITS

FIGURE 4-8

Conceptual Representation of the "Smearing" Problem
Associated With Using Oil as a Factor
(Scale is Arbitrary)

is essentially one of determining the appropriate spatial scale of subsequent experiments, i.e., the size of quadrats and the distance between them.

A difficulty which arises is that as the distance between quadrats increases, we expect the unaccounted variability (σ_e) in abundance measures to increase, thereby reducing the power of hypotheses tests about impacts of oil. The problem is conceptually depicted in Figure 4-8. To avoid contamination we want the quadrats to be as far apart as possible, but to reduce spatial variability we want the quadrats to be as close as possible. Note that spatial variability may stabilize at some distance and not increase further as distance between units increases. The problem we have defined can be stated as follows:

> *We require a survey sampling study which will estimate the spatial scale of oil smearing as a function of the area and amount of initial oiling.*

Because contamination between units must be avoided, we want to estimate the maximum value of the "smearing" scale (defined in Figure 4-8 as d).

The following simple survey study can be conducted to determine d:

1) A specified amount of oil is applied to a given area at a fixed time of tide (low tide seems logical).

2) At each subsequent tidal cycle measurements of oil are taken at increasing distances from the point of application. These measurements determine d. Note that observations are visual, in the case of oil residues, and analytical, in the case of measuring soluble hydrocarbons in tidepools and other "residual waters".

Because we expect repetitions of this experiment to yield different values of d, a design question arises as to how many repetitions are required to obtain a satisfactory estimate of d. If we use too small a value of d we run the risk of contamination. If we overestimate d, we run the risk of unnecessarily high variability and lowered test power.

From estimation theory (see Section 3.2.2) we know that the variance ($\sigma_{\hat{d}}^2$) in our estimate of d ($\hat{d}$) from any set of replicate experiments varies inversely with the number of replicates, N:

$$\sigma_{\hat{d}}^2 = \frac{s^2}{N}$$

where, $s^2 = \frac{1}{N-1} \sum(d_i - \overline{d})$ = sample variance in d

d_i = i^{th} measurement of d

$\overline{d}$ = mean value of the measurements of d

Given an estimate of s and an acceptable value for σ_d (σ_a), we can choose the appropriate number of repetitions of the experiment, N. We have no basis for choosing values for s and σ_a at this time. Preliminary field studies are needed to provide some order-of-magnitude estimates. These preliminary studies are part of the investigator's "fiddling" process.

We want to ensure that the design value of d, call it d*, is sufficiently large that there is a low probability (say β) of it being exceeded in the impact experiments. If we assume that the estimate $\hat{d}$ is normally distributed about the true value d (a reasonable assumption by virtue of the Central Limit Theorem and the essential symmetry of the measurement process), then the value of d* giving an exceedence probability of $\beta = .001$ is:

$$d^* = \hat{d} + 3\sigma_{\hat{d}}$$

This value provides a conservative lower limit on spacing between oiled and unoiled (control) quadrats.

4.4.4 Oil Impact Experiments

Impact experiments are intended to answer the following question: Is there any consistent short-term effect due to petroleum disturbances observable in rocky shore communities over a range of environmental and petroleum conditions? Our principal concern in these experiments is to determine the magnitude and extent of effects attributable to various amounts and compositions of petroleum substances and the interaction of these effects with natural disturbances. Can these oil effects be measured and differentiated from natural disturbances? We must answer this question before we can sensibly investigate mechanisms governing effects and recovery from such perturbations.

It is evident from the preceding that design of impact experiments depend on the outcomes from feasibility studies. Because results from feasibility studies are not presently available the design which follows is somewhat unrealistic. Our general layout and approach to the experiment are of primary interest. Results from the kinds of studies and experiments we have described in preceding sections are needed to develop realistic detailed designs for impact experiments. For illustrative purposes we assume that feasible experimental methods exist and that Figure 4-8 depicts the general behavior of oil applied to an experimental unit in a rocky intertidal habitat.

Experimental Objectives

The ultimate goal of impact experiments is to test a comprehensive hypothesis such as:

> *There is no significant change in the abundance of species X within Y days due to the application of Z liters/meter*^{2} *of substance W at time T to experimental units of area A located at tidal elevation H and subject to wave exposure E.*

It is unrealistic to expect to design, much less conduct, such an experiment given the inadequate knowledge we presently have. As a step towards this goal we suggest a relatively limited experiment, which is intended to test an overall approach to oil impact experiments, as well as provide additional information about the effects of oil on intertidal systems. In particular we need to find an experimental layout, which allows us to deal with the "smearing" of oil by tides and other forces. Therefore, for the present we eliminate environmental factors such as tidal elevation and wave exposure as treatment factors in the experiment and work with a single site and elevation.

Our focus is on whether or not oiled units can be differentiated from unoiled units, without immediate concern as to interactions with environmental factors, or elaborate resolution of differences due to a variety of amounts and compositions of petroleum substances.

Therefore, at a given experimental site we seek to test the following hypothesis:

> *There is no significant change in abundance of selected intertidal species in experimental units after Y days due to the application of a known quantity and composition of oil.*

Factors and Treatments

At first thought an obvious factor in this experiment is oil and we could simply work with two levels: present/absent. (We assume that the amount and composition of oil "present" is measured and known). On the face of it, oil is the only factor in the experiment. However, the difficulty which arises is the likelihood that due to "smearing" unoiled units ("controls") must be located a relatively large distance from the oiled units, thereby introducing effects due to spatial variability. [In the extreme, imagine "control" units being located at an entirely different station than oiled units. For example, if oiled units are at Zaikof Bay, we can imagine "control" units being located at one of the other rocky intertidal stations sampled by Zimmerman and Merrell (1976)]. Spatial variability then becomes an experimental factor, but is completely confounded with the factor oil. An alternative approach introduces time rather than space as a factor completely confounded with oil. Note that we could compare units at the same location before and after treatment with oil as a way to detect changes due to oil. However, temporal variability may account for part or all of any observed differences and is completely confounded with oil.

Following McCaughran (1977), our solution to this problem is to introduce both time and space as factors, both at two levels. For time, we have before oiling and after oiling; for space we have the oiled site and the unoiled site. It is important to note that oil is not a treatment factor *per se*. As described below and in Section 3.3.2, we can use a factorial design to differentiate and test for possible effects of oil by assuming that the effects of time and space are independent.

Layout

Consider the following 2x2 factorial layout with two sites (α_1 and α_2) and two times (β_1 and β_2). A two-way ANOVA model for this layout is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad \begin{aligned} i &= 1, 2 \\ j &= 1, 2 \\ k &= 1, \ldots K \end{aligned}$$

where γ_{ij} is the space-time interaction term. We can represent the interactions terms for this layout by:

		TIME	
		β_1	β_2
SPACE	α_1	γ_{11}	γ_{12}
	α_2	γ_{21}	γ_{22}

The assumption that space and time effects are independent can be tested by the *interaction hypothesis:*

$$H_I: \qquad \gamma_{11} - \gamma_{21} = \gamma_{12} - \gamma_{22}$$

To test the independence we measure abundances at each site at two different times and estimate the 4 interaction effects. The interaction hypothesis (which is equivalent to $\gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$) is then tested for significance. If the interaction is insignificant, then the time and space factors may be assumed independent.

Now if we introduce oil (or any other manipulation) at one of the sites (say i=2) and distinguish β_1 and β_2 as time before (j=1) and time after (j=2) the perturbation at this site, then the interaction hypothesis provides a test for effects due to the presence of oil, because oil only appears in one of the interaction terms (i=2, j=2):

TIME

		BEFORE	AFTER
SITES	SITE 1	γ_{11}	γ_{12}
	SITE 2	γ_{21}	▨ γ_{22}

Oil Present

The above experimental design yields a complete block 2x2 factorial layout in which we are principally concerned with testing the *interaction hypothesis*, Measurements of abundance of selected species are made at each of two sites and then oil is applied at low tide to one of the sites. At specified times after the application of oil (say 1 tidal cycle, 2 tidal cycles, etc.) abundances are again measured at each site. Each sampling time "after" simply defines a different experiment relative to the "before" time. The appropriate size of experimental units and distance between sites 1 and 2 are determined during feasibility studies (Section 4.4.2). In addition to observations of species abundance we remind the reader of the importance of measuring the amount and composition of oil in any experiments involving oil. It may also be useful to measure the hydrocarbon content of tissue from representative organisms in experimental units. These measurements can help ascertain the actual exposure of organisms to hydrocarbons. (See Chapter 2 for elaboration on the need for these measurements).

Number of Replicates

The complete block factorial experiment is discussed in some detail in Section 3.3.3. Tables 3-3 is the general ANOVA table for single or multiple blocks showing the degrees of freedom and non-centrality parameter for each hypothesis of interest. The following analysis is based on application of this table and the associated power evaluation concepts to the 2x2 interaction hypothesis. For any particular design the power of F-tests for the main effects of sites and times will be substantially greater than what we compute for the interaction hypothesis. Unless otherwise specified we assume $\alpha = .05$.

For illustrative purposes we use the variances given in Table 4.4 as estimates of σ_e for this problem. The available data (Section 4.2.1) are not adequate for computing an estimate of σ_e for the model we are actually adopting in this case. Preferrably, we would have data for two different sites and estimate σ_e from a two-way ANOVA. (Such data exist for the Alaskan coast but were not available for us to use at the time of writing this report). We expect the σ_e estimates from Table 4-4 to be the right order-of-magnitude. Therefore, the results of our analysis provide some guidance as to the design of realistic experiments.

Figure 4-9 displays the power of the F-test versus Δ/σ_e for the interaction hypothesis. Curves for 2, 4, 8 and 16 replicates of the 2x2 complete block experiment are shown. We assume one block for the experiment. If each replicate is contained in a separate block, then the power is somewhat less. Note that eight replicates of this experiment requires sampling 32 quadrats, the number of quadrats in a single replicate of the 4x2x2x2 experiment discussed in Section 4.4.1.

Figure 4-9 illustrates that even with 16 replicates this experiment does not provide a high power for deltas less than σ_e.

We illustrate the detectable change as % of mean for Balanus as a function of number of replicates for different powers in Figure 4-10. We can compare this graph with Figure 4-2a to determine the relative power of these two experiments, remembering that 2 replicates of the 4x2x2x2 experiment require the same number of samples (quadrats) as 16 replicates of the 2x2 experiment. It is evident from Figure 4-10 that large numbers of replicates are needed to confidently detect small changes in the mean which are attributable to the interaction effect (oil).

Rather than display plots such as Figure 4-10 for all selected species, we generalize the results as shown in Figure 4-11. As in Figure 4-3, we plot Δ versus replicates for different values of σ_e for a power of .85 and a .05

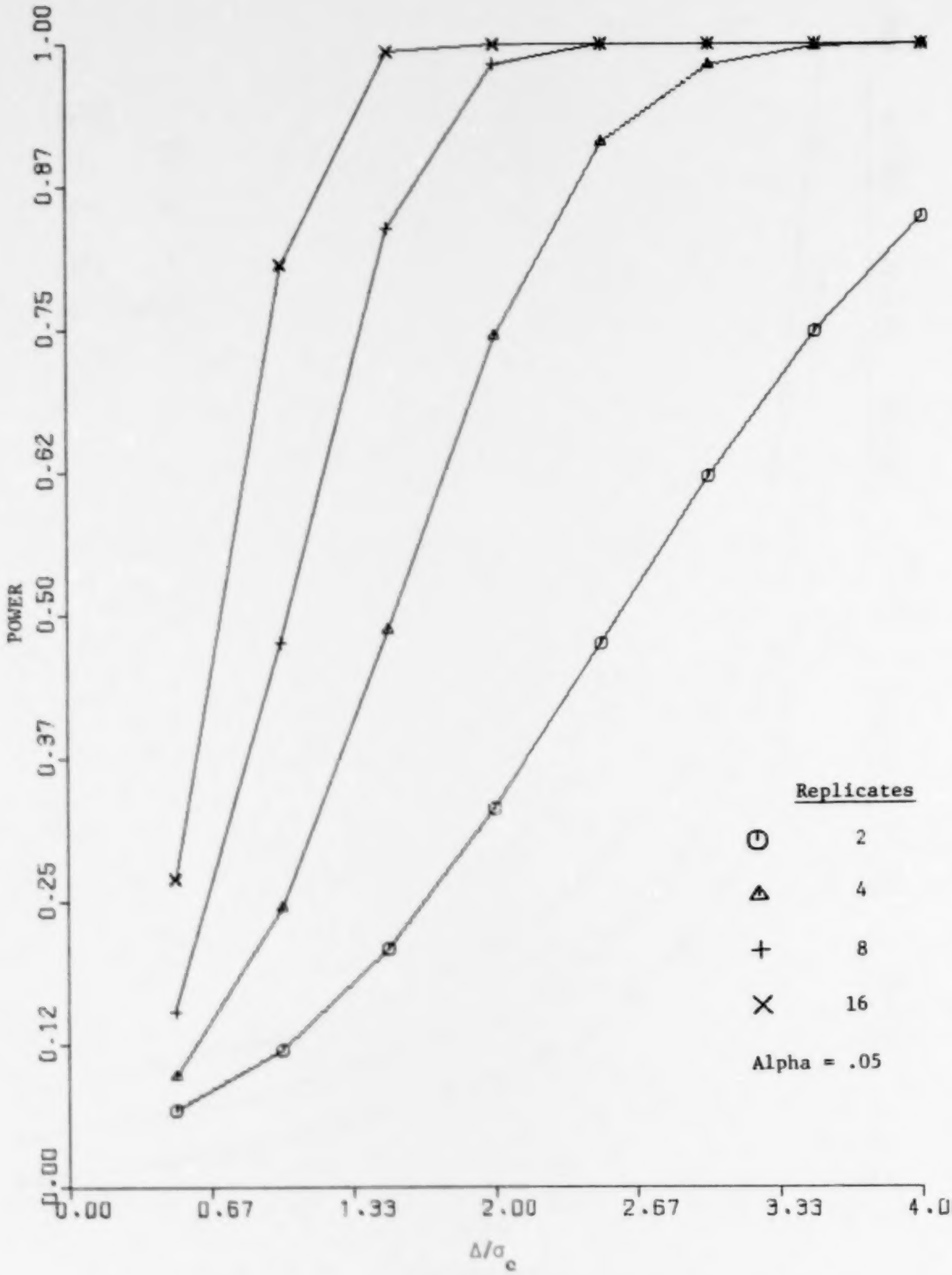


FIGURE 4-9

F-test Power vs. Δ/σ_e for Testing Interaction Hypothesis in a 2 x 2 Complete Block Layout

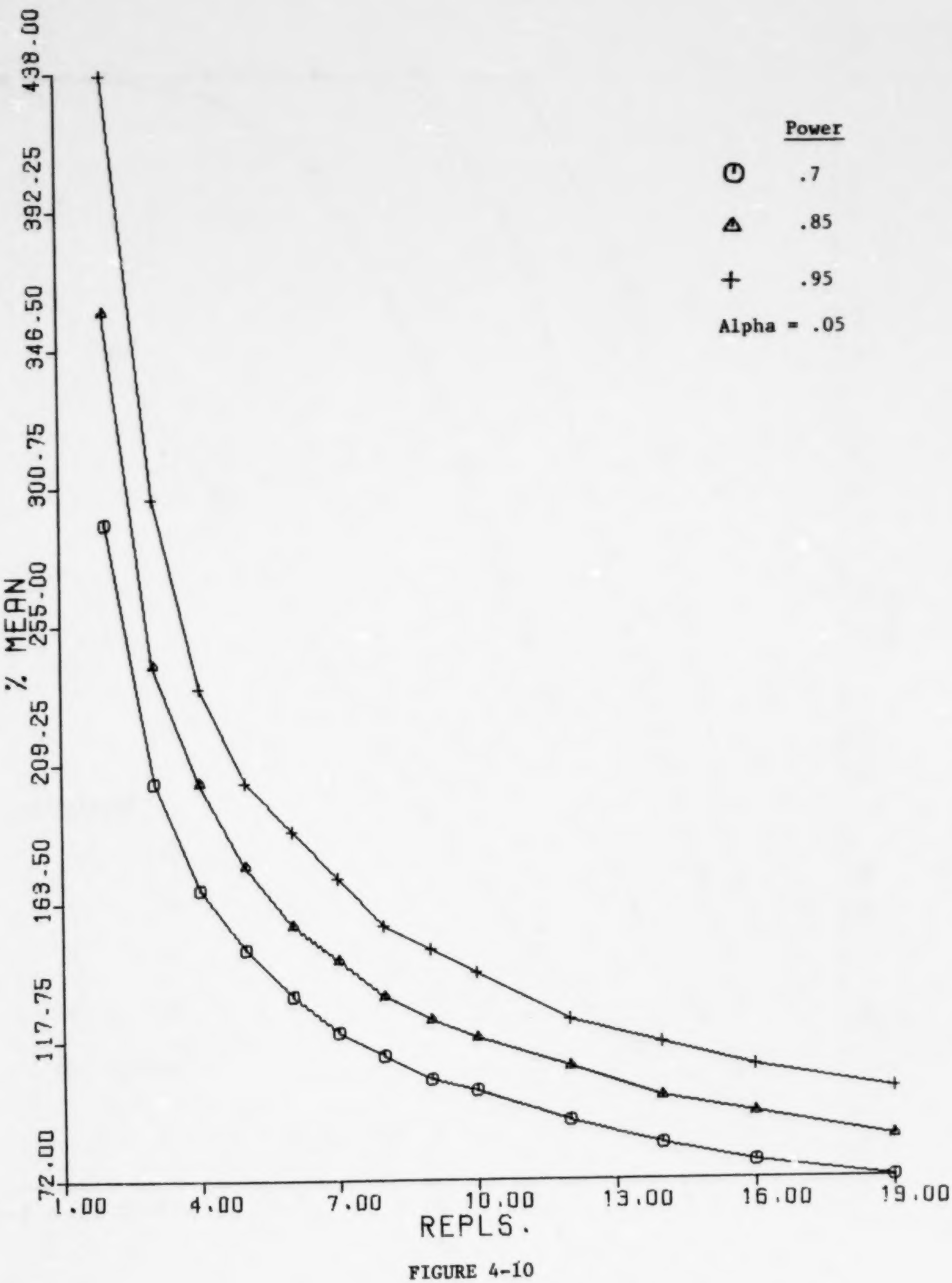


FIGURE 4-10

Detectable Percent Change in Mean of *B. glandula* Due to Interaction Effect vs. Replicates for Different Levels of Power in 2 x 2 Complete Block Layout

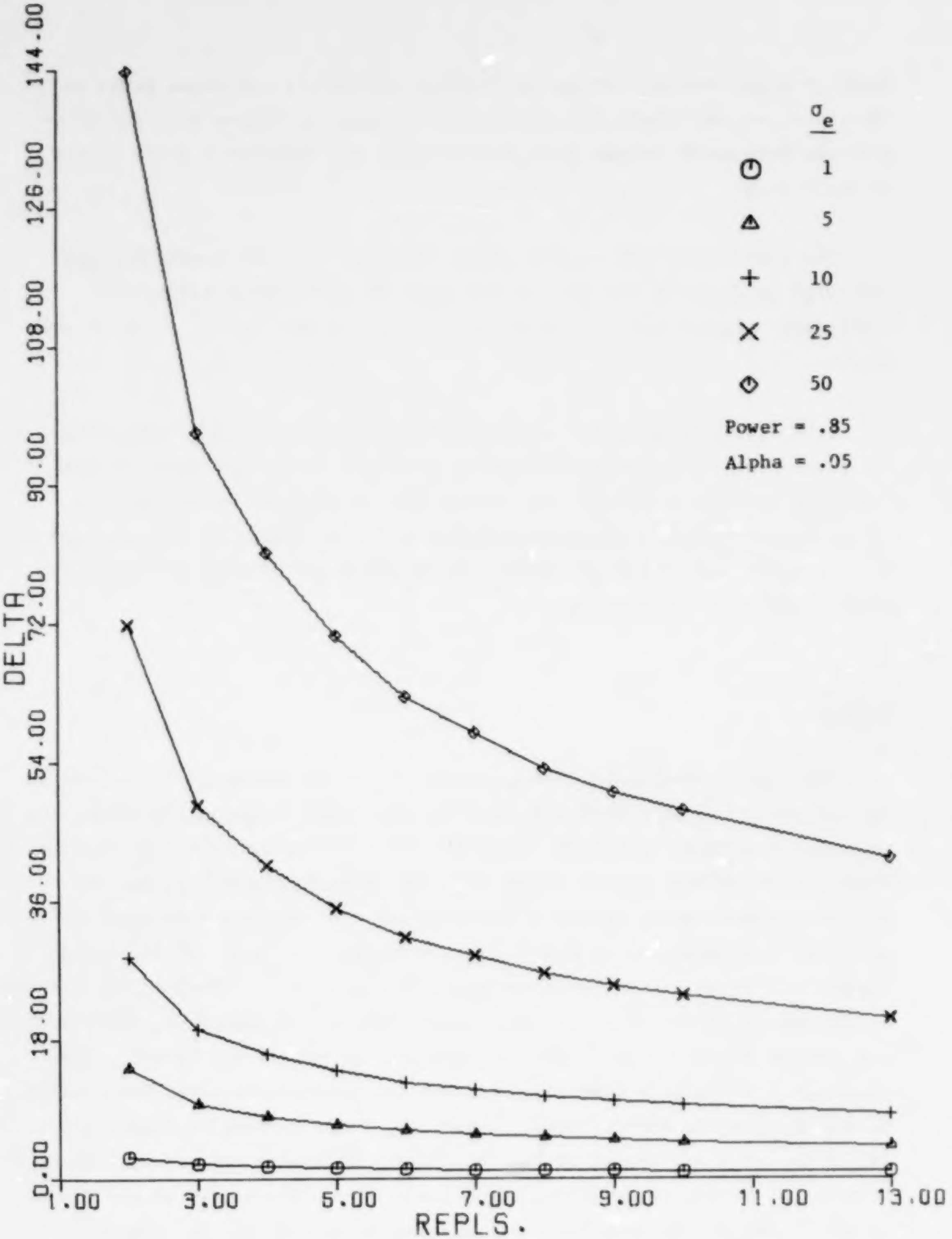


FIGURE 4-11

Detectable Change Due to Interaction vs. Replicates at Different Values of σ_e in 2 x 2 Complete Block Layout.

level of significance. Using the standard deviations and means given in Table 4-4, we can select the appropriate σ_e curve in Figure 4-11 and determine the detectable change as a percentage of the mean for a given number of replicates.

The risk trade-offs between power and alpha for this experiment are displayed in Figure 4-12a for 8 replicates and in Figure 4-12b for 40 replicates. Curves for Δ/σ_e equal to .1, .5, 1.0 and 2.0 are shown in each figure.

Finally, we display the marginal sensitivities of σ_e for increasing replicates in this experimental layout in Figure 4-13. For small deltas, this relationship is linear, indicating that an increase in the numbers of replicates yields a constant increase in the allowable σ_e for a given Δ. For very large Δ and low replicates the marginal gain in σ_e for additional replicates is greater.

Summary

Oil impact experiments raise several difficult design questions due to the nature of oil as a treatment factor. As a step towards elucidating the response of a rocky intertidal community to different petroleum substances, we suggest a relatively simple design which is as much intended to develop a feasible layout, as to measure effects of oil. A critical problem is the potential contamination of experimental controls with oil. We propose an approach in which oil is not a treatment factor *per se*. However, the effects of oil are in theory differentiable, under testable assumptions, from temporal and spatial effects. We illustrate this concept with a 2x2 layout. More elaborate layouts with additional factors can be explored after the utility of the approach is demonstrated. Relatively large numbers of replicates (8, 16 or more) are needed in this experiment to have a relatively high probability (.75 - .95) of detecting significant (α = .05) changes in the range of 50% - 100% of the mean abundance of sessile species in the community.

Note that inferences about oil drawn from an experiment such as we propose are weaker than inferences drawn from experiments in

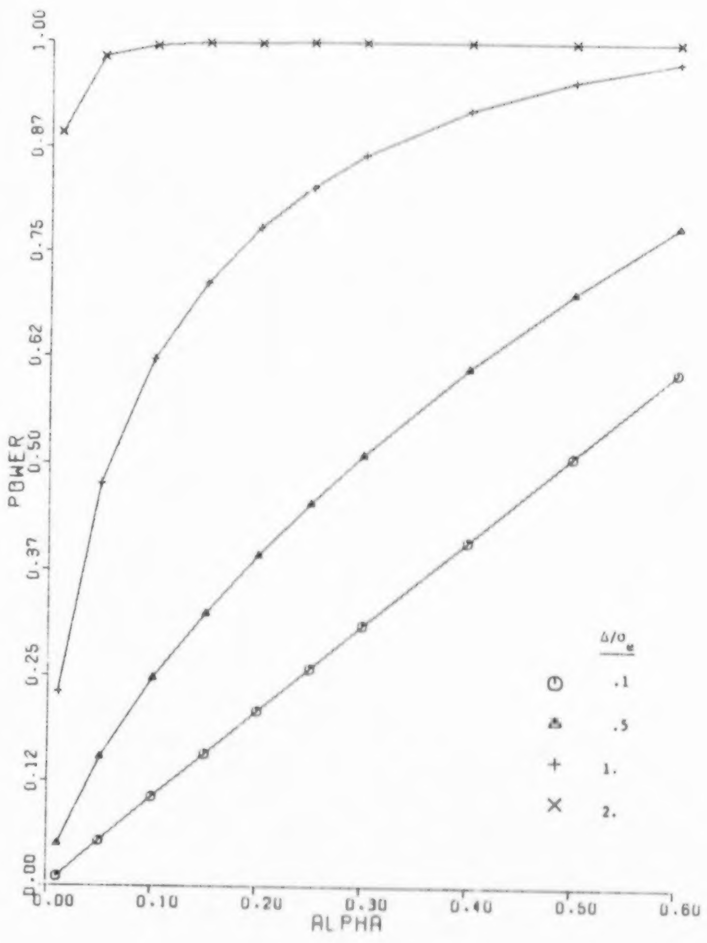


FIGURE 4-12a

Power vs. Alpha for Different Levels of Δ/σ_e with Eight Replicates of 2 x 2 Complete Block Experiment

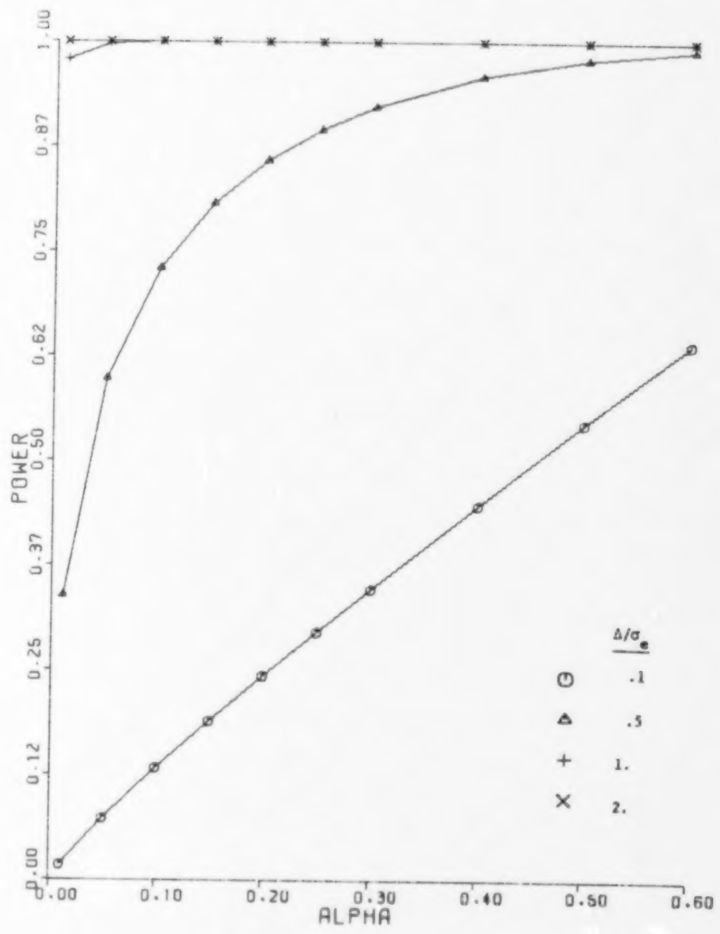


FIGURE 4-12b

Power vs. Alpha for Different Levels of Δ/σ_e with Forty Replicates of 2 x 2 Complete Block Experiment

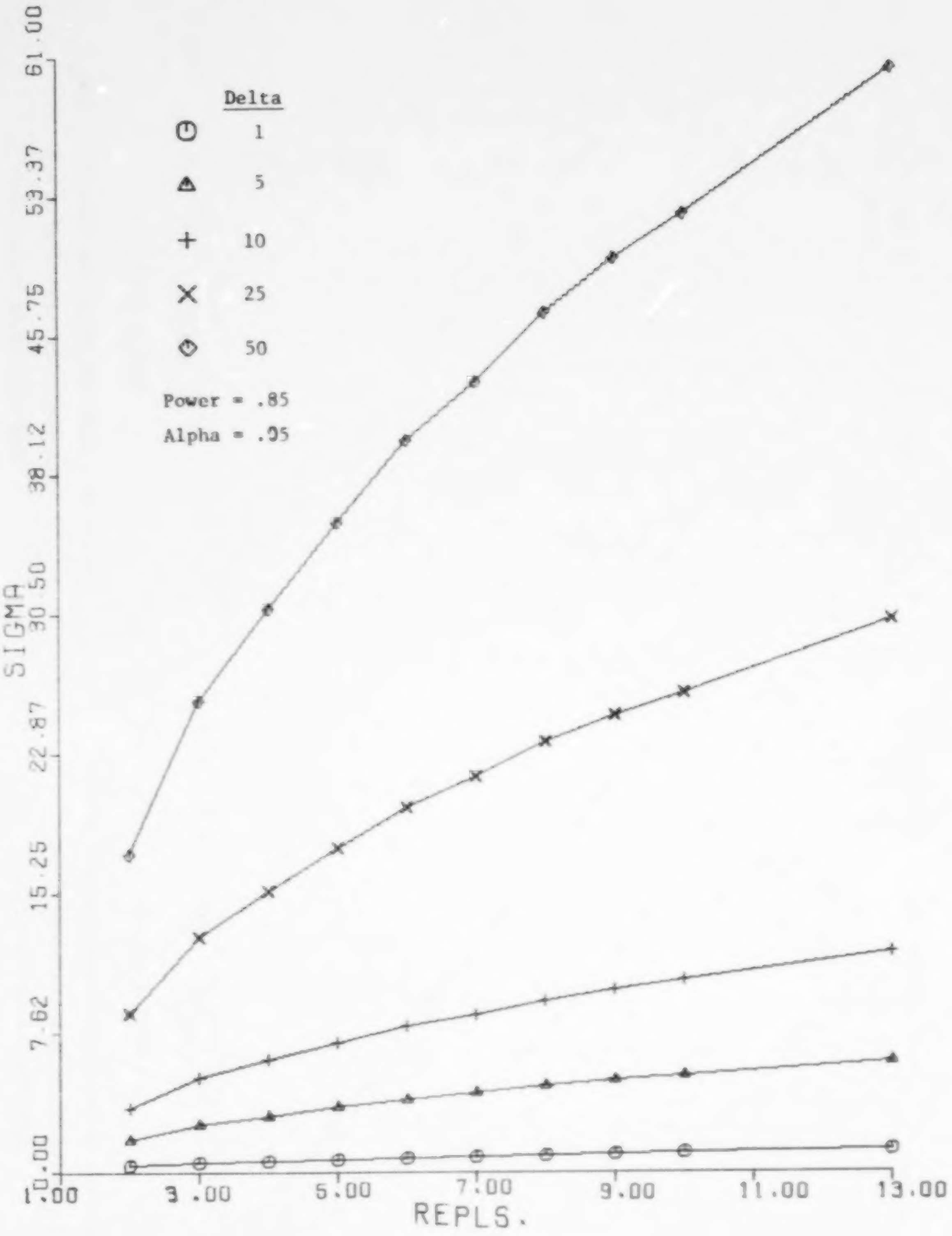


FIGURE 4-13

Marginal Sensitivity of σ_e to Replicates at Different Levels of Change for 2 x 2 Complete Block Experiment.

which oil is a treatment factor. In the design we have described oil is always confounded with temporal and spatial variability. By assuming independence of time and space effects we attribute certain changes represented by the interaction hypothesis to oil. The assumption of independence can be tested in auxiliary experiments, but not directly within the oil experiment itself.

4.4.5 Recovery Experiments

Recovery is a long-term process by which patterns of species abundance and distribution return to "normal" following an oil-caused perturbation. Included in recovery are any effects of oil manifested on a time scale beyond the initial impacts. Recovery is particularly dependent on settlement and recruitment of space by various sessile species. Of course, recovery experiments only make sense if the impact experiments have demonstrated a measurable effect due to oil. The design of recovery experiments will depend on the specific results of the impact experiments--species affected, degree of effect and so on.

We have too little information and experience to meaningfully design a recovery experiment at this time. In addition to needing more information about the feasibility of oil experiments and impacts of oil, there are many uncertainties about recovery processes, and even the meaning of recovery, in rocky intertidal habitats. If oil experiments prove feasible in the field, then at some future date, pursuit of recovery experiments can be undertaken.

4.5 IMPLICATIONS AND GENERALIZATIONS

In the following paragraphs we highlight some underlying generalizations that emerge from looking at the particular experiments and design process we use.

4.5.1 Objectives, Factors and Treatments

As we have emphasized in several places in this report, experimental design involves both subjective and objective choices. Our primary concern has been to investigate the role of objective, quantitative design methods. The examples discussed in this section have heightened our awareness of the importance of the subjective judgments which must be made. Statistical methods provide little assistance in the choice of experimental objectives, factors, treatments, and hypotheses. However, given specific experimental hypotheses, quantitative experimental design methods are a useful tool for exploring cost-information trade-offs among alternative designs.

An essential element of any experimental design is the statement of experimental objectives. We have formulated our objectives in terms of hypotheses regarding effects on particular species of certain factors and/or treatments. Much confusion can be avoided by developing statements of objectives in specific concrete, observable terms.

We have distinguished between environmental factors such as tidal elevation and experimentally imposed factors, such as oil. Given an experimental objective, the choice of factors is largely dependent on the investigator's understanding of the physical, chemical and biological processes governing the system under study. Of course, some factors are explicit in the statement of objectives. Choosing important environmental factors to be included may be more difficult. Wave exposure, desiccation, tidal elevation, substrate characteristics, temperature, solar radiation, topography, orientation, slope, predation, competition, natural death are all possible factors to be included in an experiment. The factors to be included in a particular experiment must be determined by the investigator and should reflect understanding of the experimental ecosystem. Experimental factors cannot be selected by statistical methods alone. One statistical technique, randomization, is an important and useful technique for eliminating systematic bias from factors not included in the structure of the experiment. Randomization allows the investigator to exclude factors, even though they may be known to be influencing the behavior of the system.

The investigator's experience plays a further role in the design of devices (units) and techniques for accomplishing experimental manipulations and observations. We have relied on methods described by Dayton (1971), Connell (1961) and Paine (1974) for our example experiments. However, these are only illustrative. Much ingenuity and preliminary trial-and-error experimentation may be required before suitable methods, techniques and devices have been developed. Again, statistical methods alone are not sufficient in this essential phase of experimentation.

In Chapter 2 we pointed out the difficulties of using oil as a treatment. Our examples and ability to design effective experiments are limited by the difficulties presented by oil as a treatment. Frankly, we remain in doubt as to the efficacy of using oil as a treatment in the field. Crapp (1971) describes controlled field experiments in which he applied various amounts and types of oil to quadrats on rocky shore. His experience suggests that difficulties can arise with containment of the oil and contamination of unoiled units. However, his experiments did not include careful observations of hydrocarbon compositions and amounts; so, no conclusive statements can be made. His reported results suggest that valid outcomes are possible. There is a need for further investigation and experience with oil as a treatment in field experiments before definitive conclusions can be drawn.

Finally, the one overriding criteria applicable to all of these choices is to minimize the residual error variance. Virtually all of the methods and devices for experimental design which we have discussed are based, at least in part, on the goal of reducing the residual error variance.

4.5.2 Layouts and Number of Samples

In general, we find that complete block factorial layouts constitute the appropriate experimental plan for the kind of field experiments we have described. The inherent spatial structure of these experiments results in relatively small clusters of experimental units in any single location. Therefore, we have little expectation that any significant reduction in the residual error variance will be achieved by further reduction of block size.

In addition, devices for block size reduction--confounding, fractional replication and split plot layouts--reduce the information return from the experiment by preventing estimation of certain effects (terms in the model).

It is important to recognize that the layout chosen determines the model for the observations. When the investigator knows which effects and interactions are of interest, he/she can make distinct decisions among particular layouts according to whether or not a given layout provides information about the effects in question. Given two or more layouts which provide the desired information, the choice between alternatives can be based on a comparison of the probability of detecting a significant change (power) for a specified level of experimental effort (cost).

It is interesting to compare the results of our power analysis with reported experience. For example, Dayton (1971) describes rocky shore experiments in which he used on the order of 10 replicate pairs of limpet exclusion cages to test for the effect of limpets on barnacle recruitment in a three factor experiment (tidal elevation, year and limpets). Our results in Section 4.4.2 for initial manipulation experiments are consistent with Dayton's experience. For the 4x2x2x2 complete block factorial experiment we described, Figures 4-2 and 4-3 illustrate that 4-8 replicates are an appropriate size for this experiment. We cannot use these results to draw more general conclusions about the number of replicates needed in these kinds of rocky shore experiments. However, it is indicative that the statistical analysis does not lead to unrealistically large experiments.

As a final comment on our example experiments, we reiterate the need for working with "small" manageable subsystems of the environment and following a sequential strategy of experimentation, in which each experiment builds on results of a previous effort. This cautious approach may restrict the range of validity of results, require a large number of individual experiments and lead to an apparently slower accumulation of information. However, given the variability of the environment and our lack of understanding of phenomena governing these variations we see no useful alternative.

5. CONCLUSIONS AND RECOMMENDATIONS

This report represents a rather lengthy effort to answer certain questions which arise in the process of gathering information to enable prediction and assessment of impacts of oil development on marine environments. In particular, the report addresses the problems of designing field experiments intended to test and demonstrate the effects of oil spills on rocky intertidal habitats. Our approach to design relies heavily on quantitative statistical methods. However, we appreciate the extent of subjective judgment which is also required and, in fact, distinguish explicitly between qualitative and quantitative aspects of experimental design. Qualitative aspects of design are primarily dependent on ecological understanding of the system under study and lead to the choice of observations, factors, units and layout procedures. Quantitative aspects of design are primarily statistical considerations and lead to choice of decision risks, numbers of treatment levels and replicates, and estimation of residual error variance. Because the number of possible experiments is virtually limitless, we focus on the process of design, rather than emphasizing particular designs. This focus leads us to the ecological and statistical design guidelines or criteria which are summarized at the end of Chapters 2 and 3 respectively. Although it is convenient for presentation purposes to treat ecological and statistical criteria separately, they are in fact closely related, as demonstrated by the example designs in Chapter 4.

5.1 CONCLUSIONS

At the conclusion of each chapter we have summarized appropriate ecological or statistical results in the form of specific design criteria. At this point we step back from these details and identify a small set of overall conclusions based on our investigation of experimental design. Readers interested in more detailed conclusions are referred to the individual chapter summaries.

1. *It is possible to design field experiments which will yield ecologically and statistically significant information about effects of oil on intertidal ecosystems.*

   The examples presented in Chapter 4 illustrate a process of experimental design which leads to ecologically and statistically meaningful designs. The first step in the design process is to formulate experimental objectives as explicit statements of hypotheses to be tested. From our knowledge and assumptions about possible causal factors in rocky shores, we then select experimental factors and treatments. Alternative layouts of experimental units (sampling quadrats) and alternative methods for assigning treatments to units are proposed. Power analysis is used to evaluate the probability of detecting a significant effect for each alternative. The detection probabilities identify which designs, if any, are statistically and logistically feasible and, in addition, provide an objective means for selecting the "best" of the available feasible designs.

2. *Traditional experimental design methods of factor selection, randomization, blocking and performance evaluation are directly relevant to intertidal oil experiments.*

   Statistical design methods have not often appeared useful in field studies. Natural patchiness, variability and complexity of intertidal ecosystems can result in infeasible designs and data which do not verify visually obvious effects. Oil is difficult to apply effectively as an experimental treatment. These problems motivate the need for flexible and innovative approaches to factor selection and quadrat placement, but they do not invalidate in any way the basic concepts provided by traditional experimental design theory.

3. *The experimental effort (cost) required to confirm that a particular ecological effect is statistically significant can be systematically evaluated as a function of experimental design variables.*

   In order to systematically and quantitatively design an experiment, an investigator must explicitly specify the magnitude of ecological effects of interest, the acceptable risks associated with experiment decision outcomes, and the variability in data not accounted for by the experimental model. Given these design specifications the sample size (cost) for a given experimental layout can be calculated. It remains for the decision maker to determine whether the expected information return warrants the expenditure of the required effort. Furthermore, the implementation of a given design does not guarantee return of the expected information. The actual data realized in a given case reflect the uncertainties estimated in the design method and these uncertainties may in fact lead to incorrect decisions.

4. *Sufficient experience and knowledge presently exist with rocky intertidal habitats and petroleum substances to devise explicit, meaningful hypotheses for field experiments.*

The system defined by petroleum substances (oil) interacting with a rocky shore is complex, variable, uncertain and difficult to observe. Relatively few controlled experiments have been conducted in rocky shores. Many characteristics of oil make it difficult to apply in the field, as an experimental treatment. In spite of these difficulties, an experimental sub-system, factors, units and observations can be chosen which allow formulation of testable hypotheses about specific population/community effects of spilled oil. Designing a manageable experiment may require working with a restricted range of variables and may therefore reduce the rate of information return from experimental investigations.

5. *The levels of variability exhibited by data collected along Alaskan rocky shores are sufficiently low that realistic experiments can be designed to study the effects of oil.*

A major difficulty often associated with ecological field experiments is the inability to differentiate significant experimental effects from other sources of variability in the data. Using Zaikof Bay, Alaska as a representative hypothetical experimental site, analysis of available data demonstrates that even with simple experimental models the residual variability is low enough to provide designs that are logistically feasible to implement. In addition, these designs are consistent with the size of field experiments previously reported in the literature for other rocky shore sites.

6. *A sequential design strategy which includes sample surveys and controlled manipulation experiments provides an overall approach to field experimentation most likely to maximize the long-run information return.*

Five discrete phases define a more or less chronological sequence of investigations: 1) preliminary surveys; 2) initial manipulation experiments; 3) oil feasibility experiments or surveys; 4) oil impact experiments; and 5) recovery experiments or surveys. Each stage in this sequence is less well-defined than its predecessor and at the present time more difficult to design. The results of any particular study could give rise to a myriad of alternate directions for further investigation. However, the general process of design developed and illustrated in this report is applicable to each stage in the sequence, whether it is a survey study or a controlled experiment.

5.2 RECOMMENDATIONS

Several areas for further investigation emerge from the results of this study. The following recommendations include suggestions of implementing field experiments, further analysis of existing data, development of design methods for accidental spills, and theoretical investigations of statistical design problems.

1. *Implement the initial stages of a sequential experiment design strategy at a selected study site.*

   A team of investigators including field ecologists familiar with the experimental site and statisticians familar with experimental design technology, should be funded to implement specific designs for investigating the effects of oil in field studies. Careful attention should be devoted to refining, improving and verifying the experimental design process, as well as investigating specified experimental hypotheses. The utility of the design process described in this report can only be tested by implementation of experiments and comparison of the design with actual results. It is important to note that many of our design concepts can and should be tested in the field even if controlled application of oil is unfeasible either for logistic or policy reasons.

2. *Further analyze existing NOAA data sets to investigate alternative experimental models.*

   The data analysis illustrated in this report is restricted to a small sub-set of available data and to a few simple experimental models which were sufficient for the purposes at hand. Given the relative success of this initial analysis more elaborate investigations are warranted. Specifically, further analysis should:

   a) refine the treatment of spatial variability

   b) address temporal (seasonal) variability

   c) test the assumption made in Section 4.4.4 of independence of time and space as factors

   d) include sites other than Zaikof Bay, and

   e) investigate levels of taxonomic complexity other than the species level used throughout this report.

All of these refinements are straightforward extensions of the analysis described in Chapter 4.

3. *Develop sample survey designs and response strategies for scientific study of accidental oil spills.*

   The problem addressed throughout this report is the design of controlled experiments to study oil. Accidental oil spills also occur and provide opportunities for collecting data which reveal additional information about cause and effect relationships governing the response of marine environments to spilled oil. Because accidental oil spills are neither controlled, nor replicated, sampling design methods other than those used in this report may be required. Using existing data and the basic statistical concepts of estimation and hypothesis testing, the methods developed in Chapters 3 and 4 should be extended to include the accidental spill design problem.

4. *Extend the design methodology to include random effects models and multivariate methods.*

   Since both random effects and multivariate models are commonly used in the analysis of experimental results, it seems appropriate to consider in more detail their role in experimental design. The difficulties with random effects and multivariate procedures which were briefly mentioned in Chapter 3 and Appendix A are by no means insurmountable, particularly with the widespread adoption of computerized methods for evaluating power functions and multivariate covariance matrices. A more thorough review of these methods will establish their practical relevance to the design of field experiments.

REFERENCES

Anderson, T. W., 1958, An Introduction to Multivariate Statistical Analysis, John Wiley, New York.

Blumer, M., Ehrhardt, M., and J. H. Jones, 1973, "The environmental fate of stranded crude oil," Deep-Sea Res., 20: 239-259.

Boesch, D. F., Hershner, C. H., and J. H. Milgram, 1974, Oil spills and the marine environment, a report to the energy policy project of the Ford Foundation, Ballinger Publishing Co.

Chan, G. L., 1973, A Study of the Effects of the San Francisco Oil Spill on Marine Organisms, proceedings of Joint Conference on Prevention and Control of Oil Spills, March 13-15, 1973, Washington, D.C.

Clark, R. C., Jr., Finley, D. S., Patten, B. G., Stegani, D. F. and E. E. DeNike, 1973, "Interagency investigations of a persistent oil spill on the Washington Coast: animal population studies, hydrocarbon uptake by marine organisms, and algae response following the grounding of the troopship General M. C. Meigs," proceedings of Joint Conference on Prevention and Control of Oil Spills, March 13-15, 1973, Washington, D.C., American Petroleum Institute.

Clark, R. C., Jr., Finley, J. S., Patten, B. G., and E. E. DeNike, 1975, "Long-Term Chemical and Biological Effects of a Persistent Oil Spill Following the Grounding of the General M. C. Meigs," proceedings of Joint Conference on Prevention and Control of Oil Spills, Washington, D.C., American Petroleum Institute.

Cochran, W. G., 1953, Sampling Techniques, John Wiley & Sons, Inc., Canada.

Cochran, W. G., and G. M. Cox, 1957, Experimental Designs, John Wiley & Sons, Inc., Canada.

Connell, J. H., 1961, "The Influence of Interspecific Competition and Other Factors on the Distribution of the Barnacle *Chthamalus stellatus*," Ecology 42, 710-723.

Connell, J. H., 1970, "A predator-prey system in the Marine Intertidal Region. I. *Balanus glandula* and Several Predatory Species of *Thais*," Ecol. Monogr. 40, 49-78.

Connell, J. H., 1972, "Community Interactions on Marine Rocky Intertidal Shores," Ann. Rev. Ecol. and Syst. 31, 169-192.

Connell, J. H., 1973, in Mariscal R., Experimental Techniques in Ecology, Academic Press.

Cox, D. R., 1958, Planning of Experiments, John Wiley & Sons, Inc., Canada.

Crapp, G. B., 1971, "Field Experiments With Oil and Emulsifiers," in Cowell, E. B. (ed), The Ecological Effects of Oil Pollution on Littoral Ecommunities, Elswier Publishing Co., Amsterdam.

Dayton, P. K., 1971, "Competition, Disturbance and Community Organization: The Provision and Subsequent Utilization of Space in a Rocky Intertidal Community," Ecol. Mon., 41(4), 351-389.

Dayton, P. K., 1975, "Experimental Evaluation of Ecological Dominance in a Rocky Intertidal Algal Community," Ecol. Mon., 45(2), 137-159.

Draper, N. R., and H. Smith, 1966, Applied Regression Analysis, John Wiley & Sons, Inc.

Elliott, J. M., 1971, Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates, Freshwater Biological Association, Scientific Publication No. 25.

Federov, V. V., 1972, Theory of Optimal Experiments, Academic Press, New York.

Fisher, R. A., 1926, The arrangement of field experiments. J. Ministry Agriculture Vol. 33, pp. 503-513, included as paper 17 in Contributions to Mathematical Statistics by R. A. Fisher, John Wiley, New York.

Fisher, R. A., 1935, The Design of Experiments, Oliver and Boyd, Edinburgh.

Gilfillan, E. S., Mayo, D., Hanson, S., Donovan, D., and L. C. Jiang, 1976, "Reduction in Carbon Flux in Mya arenaria caused by a spill of No. 6 Fuel Oil," Marine Biology 37, 115-123.

Kendall, M., and A. Stuart, 1973, The Advanced Theory of Statistics, Vol. 2., Inference and Relationships, Hafner Press, New York.

Kendall, M., and A. Stuart, 1976, The Advanced Theory of Statistics, Vol. 3., Design and Analysis, and Time Series, Hafner Press, New York.

Kozloff, E. N., 1973, Seashore Life of Puget Sound, the Strait of Georgia, and the San Juan Archipelago, University of Washington Press.

Lewis, J. R., 1964, The Ecology of Rocky Shores, English Universities Press, London.

Malins, D. C., (editor), 1977, Petroleum in Marine Environments and Organisms Indigenous to the Arctic and Subarctic, Academic Press.

McCaughran, D. A., 1977, "Quality of Inferences Concerning the Effects of Nuclear Power Plants on the Environment," proceedings of the Conference on Assessing the Effects of Power-Plant-Induced Mortality on Fish Populations, edited by W. Van Winkle, Oak Ridge National Lab, Oak Ridge, Tenn.

Michael, A. D., Van Raalte, C., and L. S. Brown, 1975, Long Term Effects of an Oil Spill at West Falmouth, In: Proc. API-EPA Conf. Prevention and Control of Oil Pollution, San Francisco, 1975, American Petroleum Institute, Washington, D.C.

Moore, S. F., and R. L. Dwyer, 1974, "Effects of Oil on Marine Organisms: A Critical Assessment of Published Data," Water Research, 8: 819-827.

National Academy of Sciences, 1975, Petroleum in the Marine Environment, Workshop on Inputs, Fates and Effects of Petroleum in the Marine Environment, Airlie, Virginia, May 21-25.

Nelson-Smith, A., 1972, Oil Pollution and Marine Ecology, 260 pp., Elek Science, London.

Nicholson, N. S., and R. L. Climberg, 1971, The Santa Barbara Oil Spills of 1969: A Post-Spill Survey of the Rocky Intertidal, pp. 325-399. In: D. Straughan, editor, Biological and Oceanographical Survey of the Santa Barbara Channel Oil Spill 1969-1970, Vol. 1. Allan Hancock Foundation, University of Southern California.

North, W. J., 1964, Successive Biological Changes Observed in a Marine Cove Exposed to a Large Spillage of Mineral Oil, Comm. Int. Explor. Sci. Mer. Medit. Symp. Pullut. Mar. par Micoorgan. Prod. Petrol., Monaco.

Paine, R. T., 1966, "Food Web Complexity and Species Diversity," American Naturalist 100, 65-75.

Paine, R. T., 1969, "The Pisaster-Tegula Interaction: Prey Patches, Predator Food Preference and Intertidal Community Structure," Ecology 50, 950-961.

Paine, R. T., 1971, "A Short-Term Experimental Investigation of Resource Partitioning in a New Zealand Rocky Intertidal Habitat," Ecology 52, 1096-1106.

Paine, R. T., 1974, "Intertidal Community Structure," Oecologia 15, 93-120.

Ranwell, D. S., 1968, "Extent of Damage to Coastal Habitats Due to the Torrey Canyon Incident," in Carthy, J. D. and D. R. Arthur (eds.) The Biological Effects of Oil Pollution on Littoral Communities, supplement to Volume 2 of Field Studies, Field Studies Council, London.

Rao, C. R., 1952, Advanced Statistical Methods in Biometric Research, John Wiley, New York.

Ricketts, E. F., Calvin, J., and J. W. Hedgpeth, 1968, Between Pacific Tides, Stanford California, Stanford University Press.

Scheffe, H., 1959, The Analysis of Variance, John Wiley, New York.

Smith, J. E. (ed.), 1968, Torrey Canyon Pollution and Marine Life, Cambridge University Press.

Smith, W., 1973, "An Oil Spill Sampling Strategy," unpublished manuscript, Woods Hole Oceanographic Institute, Woods Hole, Mass.

Straughan, D., 1971, Biological and Oceanographical Survey of the Santa Barbara Channel Oil Spill, 1969-70, Vol I: Biology and Bacteriology; Vol II: Physical, Chemical and Geological Studies, Allan Hancock Foundation, University of Southern California.

Thomas, M. L., 1971, "Effects of Bunker C Oil on Intertidal and Lagoonal Biota in Chedabucto Bay, Nova Scotia," J. Fish. Res. Board Can. 30, 83-90.

Wolfe, D. A. (editor), 1977, Fate and Effects of Petroleum Hydrocarbons in Marine Ecosystems and Organisms, proceedings of a Symposium, Nov. 10-12, 1976, Seattle, Washington, Pergamon Press.

Zimmerman, S. T., and T. R. Merrell, 1976, Baseline Characterizations, Littoral Biota, Gulf of Alaska and Bering Sea, final report for Environmental Assessment of the Alaskan Continental Shelf, Environmental Research Laboratories, Boulder, Co.

Zimmerman, S. T., et al, 1977, Baseline/Reconnaissance Characterization, Littoral Biota, Gulf of Alaska and Bering Sea, submitted as part of the Final Report, Outer Continental Shelf Energy Assessment Program, U.S. Dept. of Interior, Bureau of Land Management.

APPENDIX A

ESTIMATION AND HYPOTHESIS TESTING WITH MIXED EFFECT MODELS

This Appendix briefly considers how the estimation and hypothesis testing concepts outlined in Section 3.2 may be extended to models which contain both fixed and random effects. Both types of effects can be explicitly accounted for in the general linear model if Equation (3.2) is rewritten in the following form:

$$Y = \mu + X_F\theta + X_I u + X_R v + e \qquad \text{(A-1)}$$

where μ = N-dimensional vector with all components equal to the general mean of the observations

θ = P-dimensional vector of fixed main effects and interactions

u = S-dimensional vector of interactions between fixed and random factors

v = Q-dimensional vector of random main effects and interactions

X_F, X_I, X_R = partitioned components of the observation matrix X.

This partitioned equation can be viewed as a type of two-way model, in which the two "factors" of concern are deterministic (fixed) and random phenomena. Note that u and v are both random while θ is fixed, as before

Since random effects are usually defined to be zero mean, the variances of these effects are the parameters of primary interest. Estimation of these variances is complicated by the correlation structure of the observations, which is more complex than in the fixed effect case (Kendall and Stewart, 1973, Chapter 36). Problems also arise in some of the power evaluations, as shown by the brief discussion provided in the following paragraph.

Three types of hypotheses are of interest in mixed effect analyses:

$$H_F: \quad A\theta = c$$

$$H_I: \quad \sigma_I^2 = 0 \qquad (\text{or } \sigma_I^2 \geq a\sigma_e^2)$$

$$H_R: \quad \sigma_R^2 = 0 \qquad (\text{or } \sigma_R^2 \geq b\sigma_e^2)$$

Where σ_I^2 = s dimensional vector composed of the variances of each component of u

σ_R^2 = q dimensional vector composed of the variances of each component of v

a,b = specified vector of constants used to relate σ_I^2 and σ_R^2 to the residual error variance σ_e^2

Scheffé (1959) discusses the analysis of variance procedures required to test each of these hypotheses and provide appropriate ANOVA tables. Generally, the test statistics used in mixed effect situations are ratios of mean-squares which are not necessarily described by the F distributions of fixed effect analyses. The test statistic distributions for the mixed case are summarized in Table A-1. This table shows that power computation is more difficult for mixed models than for fixed effect models, where the central and non-central F distributions can be applied in a straightforward way to all hypotheses of interest.

Table A-1

DISTRIBUTIONS OF MIXED-EFFECT TEST STATISTICS

Type of Hypothesis	Hypothesis True	Hypothesis False
H_F	Hotelling's T^2	Non-central F
H_I	Central F	No standard distribution applies
H_R	Central F	Central F

Scheffé suggests that approximate F distributions may be adequate for many mixed effect applications where the correct distributions are either unknown or overly complicated. It is debatable whether the extra complexity associated with incorporating mixed effect models into the experimental design process is really justified if the final results are based on questionable approximation. Of course, mixed models may be useful or even essential for a proper *analysis* of experimental results. Our decision to ignore them is made only to facilitate experimental *design*. In general, we should expect to use more refined models and hypothesis-testing techniques in analysis than we use in design. This approach is consistent with the differing objectives and resources associated with the pre-experimental and post-experimental phases of research.

APPENDIX B

ALTERNATIVE TECHNIQUES FOR ASSIGNING TREATMENTS TO SAMPLING QUADRATS

Section 3.3.2 groups commonly used treatment allocation into two broad categories:

1. Randomized Block Techniques
    a. Complete Block Layouts
    b. Incomplete Block Layouts
        1) Confounded and split-plot designs
        2) Fractionally replicated designs
        3) Combinations of confounding and fractional replication

2. Other Techniques for Grouping Experimental Units
    a. Latin Squares
    b. Cross-over Designs
    c. Lattice Designs
    d. Nested Designs

This Appendix discusses in somewhat more detail those allocation techniques most relevant to biological field experiments. Much more information on any of the layouts discussed here may be found in references such as Cochran and Cox (1957) or Cox (1958).

In order to evaluate the relevance of available allocation schemes to intertidal impact studies, it is useful to note that most of the models of interest in Chapter 3 and 4 are *factorial* models designed to identify independent variables which have significant effects on abundance. We feel that randomized block techniques (Category 1 above) offer considerably more flexibility for testing interactions in such models and are, moreover, usually easier to use. Latin squares and lattice designs become quite complicated when several factors with different numbers of levels are to be allocated to the cells of the square or lattice in an incomplete fashion.

Cross-over and nested designs are useful in applications where the structure of the experiment makes it difficult for the experimenter to investigate certain factor combinations. Although such situations can occur in intertidal studies, they do not appear to be common and they can often be avoided by a redefinition of experimental objectives (see the petroleum-related example in Section 3.3.2). For these reasons, we have decided to concentrate on randomized block allocation techniques in this Appendix. Further details on latin squares, cross-over, and lattice designs are provided in Cochran and Cox (1957) and Cox (1958) and an extensive discussion on nested designs is included in Chapter 5 of Scheffe (1959).

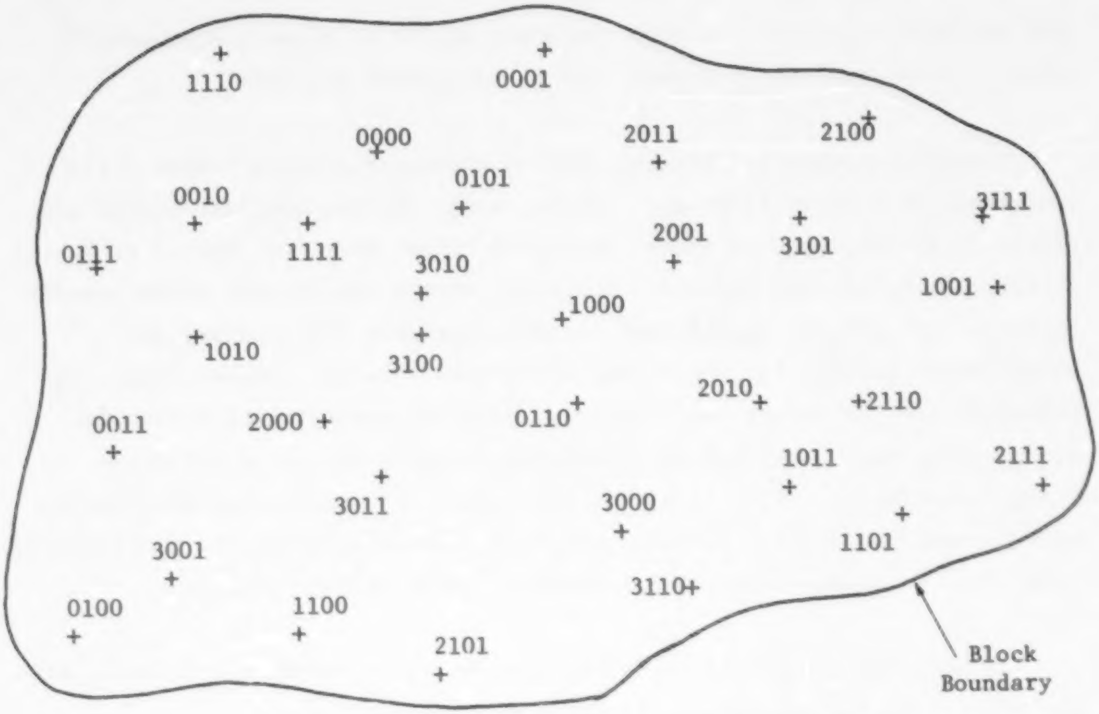
As noted in the above list, randomized block allocation techniques can be based on either complete or incomplete blocks of units. A *complete block* contains enough sampling quadrats to receive at least one complete set of factorial treatments. If, for example, a 4x2x2x2 model is being investigated a complete block would have to contain at least 32 quadrats. An *incomplete block* is any block which receives less than a complete set of treatments. As might be expected, certain main effects and/or interactions in a factorial model become difficult to estimate when the experimental blocks are incomplete. Sketches of typical complete and incomplete block layouts for the 4x2x2x2 example are shown in Figure B-1.

COMPLETE BLOCK DESIGNS

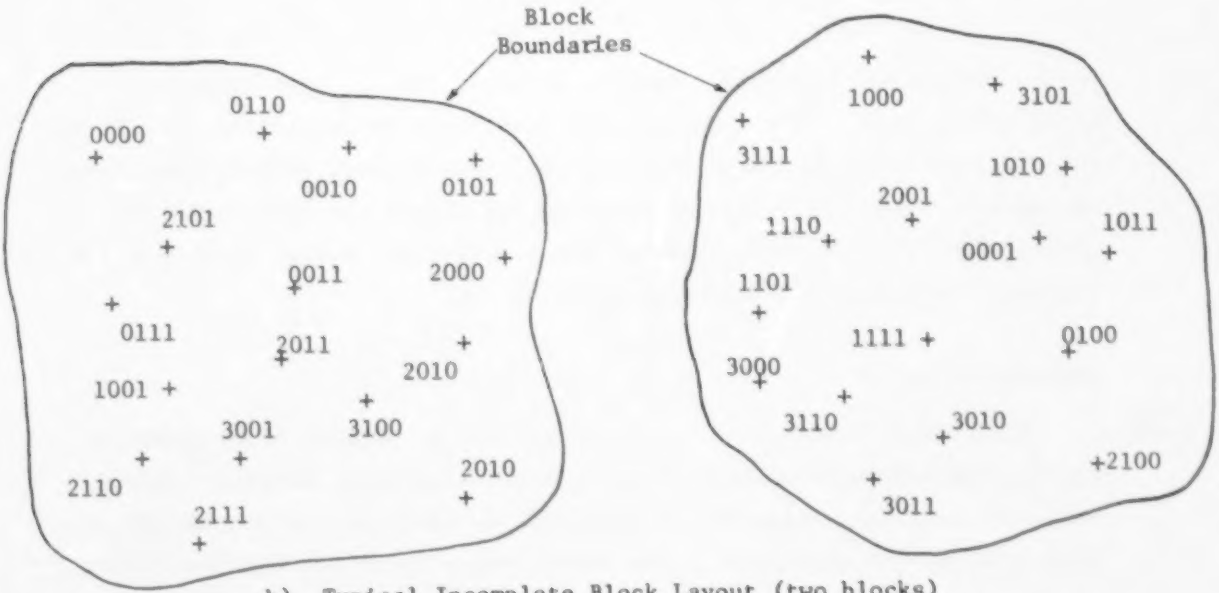
Complete block experimental designs are appealing for a number of reasons. They are easy to lay out and, consequently, usually easy to analyze. Since each factorial combination can be identified with a particular quadrat, all main effects and interactions can be unambiguously estimated if more than one replicate is available.* If several complete blocks are used in a layout,

---

**Replication can be accomplished through the use of several single-replicate complete blocks or through the use of one large complete block which contains several replicates. The first alternative is usually preferred since it is more likely to reduce the influence of nuisance variables.*


1110
0001
0000
2011
2100
0010
0101
3111
0111
1111
3010
2001
3101
1001
1000
1010
3100
2010
2110
0011
2000
0110
1011
2111
3011
3000
1101
3001
3110
0100
1100
2101
Block
Boundary

a) Typical Complete Block Layout (one block)
for One Replicate of a 4x2x2x2 Experiment


Block
Boundaries
0110
0000
0010
0101
2101
0011
2000
0111
2011
2010
1001
3001
3100
2110
2010
2111
1000
3101
3111
1010
2001
1110
0001
1011
1101
1111
0100
3000
3110
3010
+2100
3011

b) Typical Incomplete Block Layout (two blocks)
for One Replicate of a 4x2x2x2 Experiment

Figure B-1

Examples of Complete and Incomplete Block Layouts this estimation property enables the experimenter to clearly distinguish block effects from the treatment effects of primary interest.

Complete layouts do, however, have disadvantages which reduce their usefulness in some applications. First, truly uniform complete blocks are difficult to construct in intertidal applications when more than 3 or 4 factors are being investigated. The block dimensions in such cases usually approach the scale of significant nuisance variable fluctuations and within-block variability can become unacceptably large. Second, the total number of treatments (or, equivalently, sampling quadrats) in a complete block layout may be quite large, implying a costly or logistically unwieldy experiment. Often, the extensive amount of information provided by such layouts is not really needed since the investigator is not particularly interested in higher-order interactions but only in main effects.

We can, then, identify two situations in which incomplete block designs may offer a useful alternative to standard complete layouts:

1. Complete blocks are too large to be uniform with respect to elevation, exposure, predatory pressure or other potentially important nuisance variables.
2. Complete blocks require more quadrats than are either practical or desirable, considering experimental objectives and constraints.

In the first case, confounded designs probably are the best incomplete block alternative. In the second case, fractionally replicated designs are most appropriate. When both block uniformity and total quadrat number are of concern, a combination of confounding and fractional replication is usually the best solution. Each of these incomplete design techniques is discussed in one of the following sections.

CONFOUNDED DESIGNS

Confounded layouts obtain the convenience of smaller block sizes by sacrificing information about certain factorial effects (usually higher-order interactions) which are of little or no relevance to the primary objectives of the experiment. The sacrificed effects are less accurately estimated because they are inferred from comparisons of observations obtained from different blocks (confounded designs always use more than one block). This procedure tends to lump true factor effects together with spurious between-block effects due to nuisance variables. The factor effects are then said to be *confounded* or confused with block effects. Confounding is *complete* if treatment effects are estimated only from between-block comparisons and *partial* if treatment effects are estimated from both within and between-block comparisons. As might be expected, the F test power for a given main effect or interaction decreases as the effect becomes more confounded and its estimation accuracy drops. Completely confounded effects cannot be properly estimated or tested since they are indistinguishable from between-block nuisance variations. The concept of confounding is illustrated in Figure B-2 where two 8 quadrat complete blocks for a 2x2x2 experiment are split into four 4 quadrat incomplete blocks. This particular design confounds third-order interactions with block effects.

The estimation equation for the general third-order interaction term in the 2x2x2 model is given by Kendall and Stuart (1976, Chap. 35):

$$\hat{\gamma}_{ijk} = \bar{y}_{ijk} - (\bar{y}_{ij.} + \bar{y}_{i.k} + \bar{y}_{.jk})$$

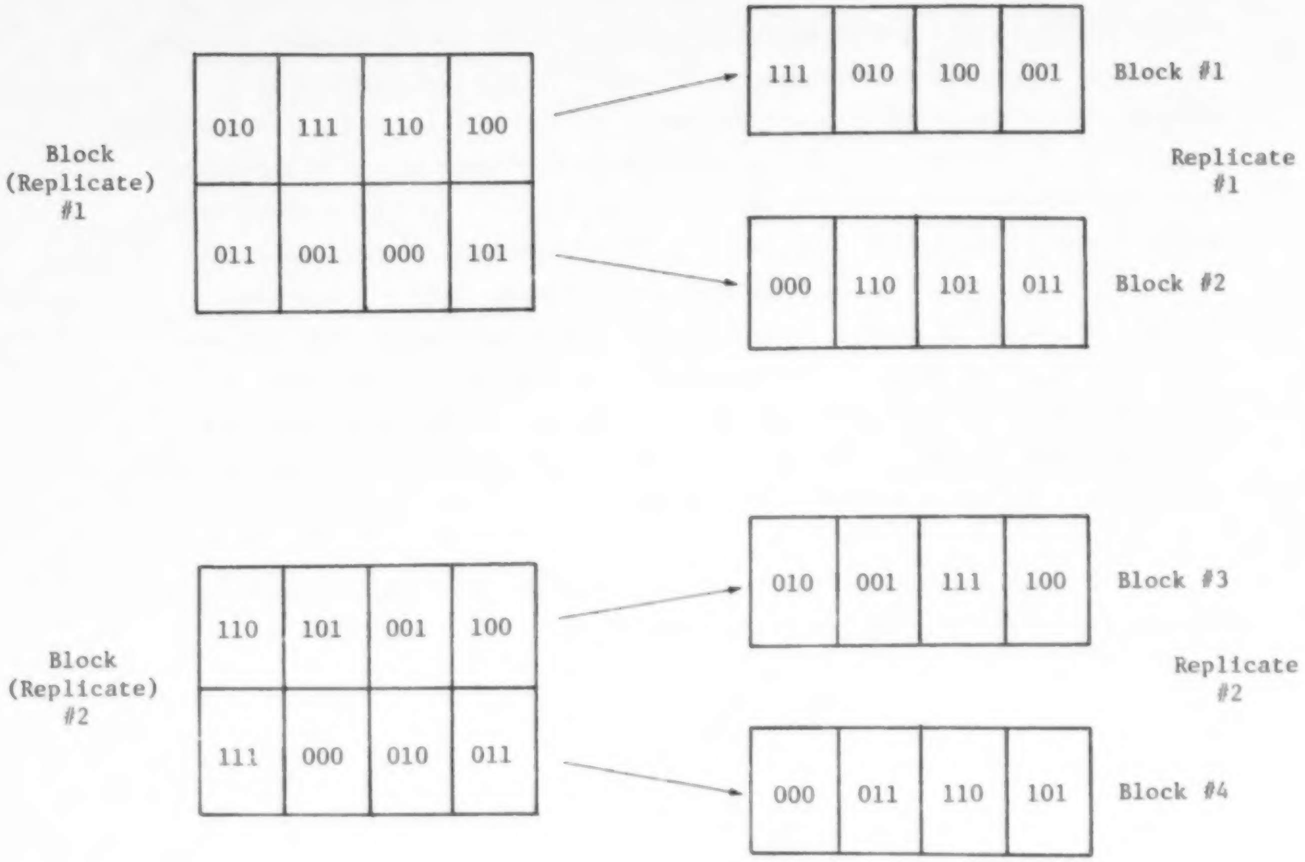
$$+ (\bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..k}) - \bar{y}_{...}$$

where y_{ijk} represents the mean of all replicates for treatment combination (i,j,k) and the dot subscripts indicate averaging over the appropriate factor. The hypothesis of no third-order interaction in the 2x2x2 model is equivalent to the requirement that $\hat{\gamma}_{ijk}$ be zero for all i,j, and k.

The ambiguity introduced by confounding is best revealed if a particular third-order effect estimate is examined. After some algebraic manipulation the estimate for γ_{111} can be written:

$$\hat{\gamma}_{111} = (\bar{y}_{111} + \bar{y}_{100} + \bar{y}_{010} + \bar{y}_{001})$$

$$- (\bar{y}_{110} + \bar{y}_{101} + \bar{y}_{011} + \bar{y}_{000})$$


Block (Replicate) #1
010
111
110
100
011
001
000
101
111
010
100
001
Block #1
Replicate #1
000
110
101
011
Block #2
Block (Replicate) #2
110
101
001
100
111
000
010
011
010
001
111
100
Block #3
Replicate #2
000
011
110
101
Block #4

a) Complete Block Layout

b) Incomplete Block Layout With Third-Order Interaction Completely Confounded

Figure B-2

Example of Confounding for a 2x2x2 Layout

This estimate is identical to the block effect estimate obtained by differencing the means of observations in each of the four quadrat blocks of Figure B-1b--i.e., the γ_{111} interaction is completely confounded with between-block variations. When the usual uniqueness constraints are applied (see Section 3.2.2) it is easily shown that *all* of the third-order interaction estimates are equal to $\pm\hat{\gamma}_{111}$. Consequently, they are all confounded with block effects and it is impossible to test the third-order interaction hypothesis. Fortunately, the lower-order interactions and main effects can be estimated unambiguously. Chapter 6 of Cochran and Cox (1957) outlines in more detail the relationship between confounded layouts and interactions involving higher-order factorial effects. The examples in this reference should be reviewed by anyone seriously interested in confounding as a design tool.

The particular effects to be confounded in an incomplete block experiment determine the general structure of the sampling layout. Although formalized procedures are available for deriving the appropriate layout, it is usually easier to rely on the standard plans provided in Cochran and Cox (1957). These plans assign treatments to the incomplete blocks so as to insure that selected effects (the most important ones) can be properly estimated (i.e., are either unconfounded or only slightly confounded). Although every level of every factor appears at least once in each block, the total number of quadrats required per block is considerably reduced (as in the above example). The treatment combinations provided in standard plans should, of course, be randomly allocated to the particular quadrats in each block.

Since main effects are nearly always important in factorial experiments, most confounded designs sacrifice higher order interactions which are of less interest. A notable exception is the *split-plot* design which completely confounds the main effect of one factor (e.g., tidal elevation) in order to obtain better estimates of other main and interaction effects. Split-plot designs offer a possible solution to the tidal elevation blocking dilemma discussed in Section 3.3.2 if the experimenter is willing to confound the main elevation effect in order to gain information about elevation--petroleum interactions. This approach is most attractive when the total number of treatments is relatively small since the experimental analysis becomes rather complicated when incomplete blocking techniques must be applied to the sub-plots within each plot. Readers interested in pursuing split-plot design methodology further are referred to Chapter 7 of Cochran and Cox (1957).

Generally, it is good design practice to partially confound several factorial effects in an experiment rather than to completely confound one or two. This more cautious approach provides *some* information on every effect in the model and does not force the experimenter to ignore interactions which may, in fact, be significant. It is easy to see that partial confounding is possible only when more than one replicate is used, since only in this case can both within-block and between-block estimates of an effect be made. The degree to which a particular effect is confounded may be conveniently measured in such situations by a ratio known as the *relative information:*

$$r = \frac{\text{Number of replicates used for within-block estimation of confounded effects}}{\text{Total number of replicates}}$$

Relative information ratios are usually provided with the plans of partially confounded layouts (see, for example, Chapter 6 of Cochran and Cox, 1957).

The general principles for the analysis of test power outlined in Section 3.3.3 can be applied directly to confounded designs, although the computations are somewhat more complicated than in the complete block case. The most straight-forward analysis is obtained when only complete confounding is used since, in this case, the confounded effects are simply deleted from the ANOVA table and not tested.

Both the non-centrality parameters and number of degrees of freedom of unconfounded effects remain unaffected by the complete confounding procedure.

The number of degrees of freedom for residual error is influenced by confounding, however, since the addition of more blocks and deletion of certain effects causes a redistribution of sampling resources. An example of this is shown in Table B-1, which summarizes the degree of freedom allocations for confounded and unconfounded I^F factorial designs with K replicates distributed equally among L_C and L_U blocks, respectively. The parameter C is the number of degrees of freedom lost to the treatment category as a result of confounding. This parameter depends on the number of confounded blocks as well as on the order of the confounded effects in all designs with $I > 2$. It is difficult to generalize about the practical influence of confounding on test power for unconfounded effects. In most cases, the influence will probably be small, particularly if several replicates are taken, but it is best to construct an ANOVA table and check the power to confirm this in each individual case.

Table B-1

COMPARISON OF DEGREE OF FREEDOM ALLOCATIONS FOR UNCONFOUNDED AND COMPLETELY CONFOUNDED DESIGNS

	Unconfounded Design	Design with Selected Effects Completely Confounded
Treatment effects	$I^F - 1$	$I^F - 1 - C$
Block effects	$L_U - 1$	$L_C - 1$
Residual error	$I^F(K-1) - (L_U-1)$	$I^F(K-1) + C - (L_C-1)$
Grad mean	1	1
Total samples	$N = KI^F$	$N = KI^F$

When treatment effects are partially confounded they are not deleted from the analysis of variance but are estimated and tested in much the same way as unconfounded effects. Their non-centrality parameters must be modified, however, to reflect the revised sum of squares computations necessitated by the confounding procedure. Each hypothesis test in a partially confounded ANOVA table should be looked upon as a separate problem, with the sum of squares taken only over the within-block comparisons available for estimation. The non-centrality parameter may then be found by replacing all observation-dependent quantities in the treatment sums of squares by their expected values (see Equation (3-17) and Scheffe, 1959, Chapter 2) and applying the "worst-case" effects assumption of Section 3.2.3.

In a partially confounded experiment, the confounded effect non-centrality parameters will be smaller than their unconfounded counterparts so, as might be expected, the F test power for these effects will be reduced. It is questionable whether or not detailed power computations for partially confounded effects are justified in practice since these effects will usually be interactions which are less important to the experimenter than main effects. If the main effects are not confounded, their power may be computed as in the complete block case, provided that the residual error degrees of freedom value is modified appropriately. Usually the major impact of confounding is not on the experiment's power functions but, rather, on the magnitude of the residual error variance. The practical benefits of confounding are, consequently, difficult to assess until experimental data have actually been collected and residual errors for confounded (small block) and unconfounded (larger block) designs can be compared. Cochran and Cox (1957, Chapter 6) give an equation for predicting (after the fact) the effect on residual error of using a complete block design instead of a confounded design. When a sequence of related experiments is being performed, such *post-facto* evaluations may help the designer gradually refine his allocation strategy. Short of this, good judgment and keen observation provide the best basis for deciding whether or not confounding is really needed.

FRACTIONALLY REPLICATED DESIGNS

Confounding provides a useful way to reduce block sizes and possibly residual errors in factorial experiments but it does not offer any reduction in the total number of samples required in each replicate. It is reasonable to ask if the large sample sizes required in even relatively modest factorial experiments are really justified in most intertidal research experiments. Most of these samples are needed to provide independent estimates for the large number of higher-order interaction effects generated by the factorial model, effects which are difficult to interpret even if they can be effectively tested. If the designer is willing to sacrifice his ability to estimate such interactions, he may use a technique known as fractional replication to obtain a substantial reduction in total sample size. When fractional replication is used in combination with confounding, an impressive amount of information can be obtained from a surprisingly small number of experimental observations.

Fractionally replicated experiments, as the name implies, retain only a specified fraction of all possible factorial treatments in each replicate. This is most easily accomplished in experiments where all factors have the same number of levels I (usually two or three). In such cases the fractions used are given by $1/I^{\alpha}$ where α is small positive integer and the treatments to be retained are determined by higher order interactions known as *defining contrasts*. The defining contrasts may be used to split the replicate into α equal parts, all equally useful for estimating main effects and interactions.

If only one of the fractional replicates is retained, the experiment will be unable to discriminate between the members of certain groups of effects called *aliases*. This is most easily shown with a simple example such as the 2x2x2 experiment. If the third-order interaction is selected as the defining contrast, the treatments are divided into the following two fractional replicates (see Chapter 6A of Cochran and Cox, 1957, for details):

Fractional replicate #1	Fractional replicate #2
100	110
010	000
001	011
111	101

Inspection of the equations for estimating the 7 main and interaction effects of the complete 2x2x2 factorial model shows that the third-order interaction cannot be estimated from either of the fractional replicates since it is completely confounded with between-replicate comparisons. The other effects can be estimated but the estimation equations are the same for the following pairs:

Main effect 1	and	Interaction effect 2,3
Main effect 2	and	Interaction effect 1,3
Main effect 3	and	Interaction effect 1,2

Each of the main effects is *aliased* with a second-order interaction. If, for example, the ANOVA test for main effect 1 is significant there is no way to know whether the significance is due to factor 1, to the interaction between factors 2 and 3, or to some combination of these. Of course, the experimenter may be willing to *assume* that higher-order aliases are negligible so that all significant results can be attributed to main effects. In some cases, this assumption may be justified by past experience or other information, in others is may just be a speculative hypothesis.

In larger factorial experiments, the designer has a fair amount of flexibility in choosing the defining contrasts for fractional replication. Generally, these contrasts should be selected so that the aliases of main effects are the interactions which are most likely to be small (usually the highest order interactions). A useful index of fractionally replicated plans is given in Chapter 6A of Cochran and Cox (1957). Once a plan is chosen, the aliases of any estimable effect may easily be found from its generalized interaction with the defining contrasts (the procedure is outlined in Cochran and Cox) or they may be obtained from the tables given in Chapter 12 of Cox (1958). In any case, it is important to remember that *all* effects estimated in a fractionally replicated experiment have aliases. Any fractionally replicated plan therefore presents the potential danger of attributing significance incorrectly to the particular effect or lower-order interaction of most interest to the researcher. So long as the experimenter appreciates this danger and accounts for it in his interpretation of experimental results, fractional replication can be a very efficient and effective allocation procedure.

It is easy to see that when fractionally replicated designs reduce the total number of samples required in an experiment they also reduce the number of samples needed in each block. If still smaller block sizes are required in order to minimize nuisance variable effects, some of the interactions in the fractional replicate can be confounded. Obviously, many alternative designs could be constructed. If, for example, two 1/4 fractional replicates were taken in a 2^F experiment, selected effects could be partially confounded, block sizes could be made as small as desired and the total sample size would still be half of a single complete replicate. The ANOVA computations and power evaluations for such sophisticated sampling layouts become rather complicated and should be carried out carefully. The savings in sampling effort obtained may, however, be well worth the extra statistical analysis required.

The significant decrease in sample size obtained from fractional replication raises some questions about the number of degrees of freedom available for estimation of residual error. To gain a broader perspective on this matter it is useful to reconsider the distribution of degrees of freedom in the I^F unconfounded complete block experiment.

Recall, from Table B-1, that the residual error degrees of freedom available in the unconfounded and completely confounded I^F designs are given by

	Unconfounded Design	Design with Selected Effects Completely Confounded
Residual error degrees of freedom	$I^F(K-1)-(L_U-1)$	$I^F(K-1)+C-(L_C-1)$

If the number of replicates (K) is small, it is possible that neither of these designs will be able to supply enough degrees of freedom to permit a residual error sum of squares to be computed. In the extreme case when only one replicate is available, the unconfounded design can have only one complete block and no degrees of freedom are available for residual error.

The number of residual error degrees of freedom in the confounded case may be very small even when more than one replicate is available since the loss in degrees of freedom due to confounding (L_C-1-C) may be significant. When this occurs, the designer has three alternatives for obtaining a residual error sum of squares value:

1. The residual error mean sum-of-squares can be computed from historical measurements taken in related experiments; provided that the residual error variance appears to be stable from experiment to experiment.
2. Some of the higher-order interactions included in the treatment sum-of-squares can be deleted from the experimental model (incorporated into the residual error term). The degrees of freedom contributed by these terms will go to residual error.
3. The number of replicates can be increased.

The first two procedures require the experimenter to make assumptions which may or may not be justifiable. Obviously, it is preferable to obtain more residual error degrees of freedom through additional replication.

The above comments apply equally well to fractionally replicated designs except that, in this case, the option of obtaining more replicates is less realistic. The primary reason for using fractional replication is to reduce sample size below the number contained in a single replicate. For this reason, most simple fractionally replicated designs use residual sum of squares obtained either from historical information or from higher order interactions which have been lumped with the error term in the experimental model. Of course, if the fractional replication is small enough (1/3 or 1/9, for example) a number of separate fractional replicates may be laid out and the total sample size still kept smaller than that contained in a complete replicate. In such cases, residual error sums of squares can be computed and, if different defining contrasts are used for different fractional replicates, a number of the lower order interaction terms may still be estimated. Further details are available in Chapter 6A of Cochran and Cox (1957).

Power analyses of simple fractional replication experiments are relatively straightforward once a method for obtaining the residual error sum of squares has been selected. The total number of degrees of freedom is, of course, reduced by a factor of I^{α} in each fractional replication. The $I^{(F-\alpha)}-1$ degrees of freedom available for treatments in each fractional replication of the unconfounded design are distributed among all estimable effects (each group of aliases shares the number of degrees of freedom allotted to one of their number in a complete block experiment). The non-centrality parameters of the aliased effects are also decreased by the I^{α} factor since the treatment sums of squares are taken over fewer observations. The combined influence of these two changes reduces test power significantly as compared with the complete block experiment. Fractional replication should, therefore, only be used when the test power for main effects is more than adequate and the complete block approach provides much more information than is really needed.

A check of the standard F test power curves will show the loss in power resulting from fractional replication in a particular experiment. For most intertidal applications, where residual errors are fairly high and the magnitude of important effects small, the sacrifice in power required by fractional replication will make very small fractional replications ($\alpha > 2$) unattractive. The 1/2 replication of the 2^F and 1/2 replication of the 3^F experiments will probably be the ones used most frequently in practice. If these do not reduce the total sample size sufficiently, the designer should seriously consider decreasing the number of factors investigated in the experiment.